Comparison of the genetic algorithm and incremental optimisation routines for a Bayesian inverse modelling based network design

Comparison of the genetic algorithm and incremental optimisation routines for a Bayesian inverse modelling based network design

A Nickless[1,2], P J Rayner[3], B Erni[1,4], and R J Scholes[5]

[1] Department of Statistical Sciences, University of Cape Town, Cape Town, 7701 South Africa

[2] Department of Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford OX2 6GG, UK

[3] School of Earth Sciences, University of Melbourne, Melbourne, VIC 3010, Australia

[4] The Centre for Statistics in Ecology, the Environment and Conservation, University of Cape Town, Cape Town, 7701 South Africa

[5] Global Change Institute, University of the Witwatersrand, Johannesburg, 2050 South Africa

E-mail: alecia.nickless@phc.ox.ac.uk

Abstract. The design of an optimal network of atmospheric monitoring stations for the observation of carbon dioxide (CO_2) concentrations can be obtained by applying an optimisation algorithm to a cost function based on minimising posterior uncertainty in the CO_2 fluxes obtained from a Bayesian inverse modelling solution. Two candidate optimisation methods assessed were the evolutionary algorithm: the Genetic Algorithm (GA), and the deterministic algorithm: the Incremental Optimisation (IO) routine.

This paper assessed the ability of the IO routine in comparison to the more computationally demanding GA routine to optimise the placement of a five-member network of CO_2 monitoring sites located in South Africa. The comparison considered the reduction in uncertainty of the overall flux estimate, the spatial similarity of solutions, and computational requirements. Although the IO routine failed to find the solution with the global maximum uncertainty reduction, the resulting solution had only fractionally lower uncertainty reduction compared with the GA, and at only a quarter of the computational resources used by the lowest specified GA algorithm. The GA solution set showed more inconsistency if the number of iterations or population size was small, and more so for a complex prior flux covariance matrix. If the GA completed with a sub-optimal solution, these solutions were similar in fitness to the best available solution.

Two additional scenarios were considered, with the objective of creating circumstances where the GA may outperform the IO. The first scenario considered an established network, where the optimisation was required to add an additional five stations to an existing five-member network. In the second scenario the optimisation was based only on the uncertainty reduction within a subregion of the domain. The GA was able to find a better solution than the IO under both scenarios, but with only a marginal improvement in the uncertainty reduction. These results suggest that the best use of resources for the network design problem would be spent in improvement of the prior estimates of the flux uncertainties rather than investing these resources in running a complex evolutionary optimisation algorithm.

The authors recommend that, if time and computational resources allow, that multiple optimisation techniques should be used as a part of a comprehensive suite of sensitivity tests when performing such an optimisation exercise. This will provide a selection of best solutions which could be ranked based on their utility and practicality.

1. Introduction

In order to understand the role of carbon dioxide (CO_2) in climate change, and to monitor mitigation efforts to reduce emissions of CO_2, estimates of land-based sources and sinks of CO_2 can be obtained through the technique of inverse modelling. This can be achieved with accurate and precise measurement of atmospheric CO_2 concentrations at suitably located monitoring sites, a reliable atmospheric transport model, and a Bayesian inverse modelling framework (Jackson 1979, Jackson & Matsu'ura 1985, Rodgers 2011, Enting 2002, Gurney et al. 2003, Tarantola 2005, Ciais et al. 2010).

The reason for choosing the Bayesian approach is due to the under-determined nature of the problem - there are far more source regions to solve for than there are measurements. Therefore prior information on the surface CO_2 fluxes is used to regularise the problem by narrowing the parameter space of the solution. Bottom-up approaches, such as land-atmosphere exchange models and fossil fuel inventory analyses, provide the required prior information on the unknown CO_2 fluxes.

Deciding on the placement of measurement sites is not always possible and instead existing measurement infrastructure may be the only source of data. But when new sites are to be installed, it is possible to exploit the experimental design in order to optimise the inversion solution for the unknown parameters (Haber et al. 2008). This paper considers the optimal experimental design for an atmospheric CO_2 monitoring network for South Africa. The network optimisation approach we have adopted, based on inverse modelling, was originally based on the work of Hardt & Scherbaum (1994) which optimised the station locations for an inversion problem applied to a seismographic network. This was developed by Rayner et al. (1996), and the approach recently reviewed in Kaminski & Rayner (2017).

Previous studies which have considered this optimisation problem for CO_2 atmospheric monitoring networks have implemented three different optimisation methods: simulated annealing (Rayner et al. 1996); incremental optimisation (IO) (Patra & Maksyutov 2002); and the genetic algorithm (GA) (Rayner 2004). Simulated annealing and the GA are evolutionary algorithms whereas the IO is a deterministic algorithm. The IO routine has already been compared with simulated annealing (Patra & Maksyutov 2002) for this application. This paper aims to compare the GA with the IO routine as described by Patra & Maksyutov (2002). Both optimisation routines are plausible candidates for the network design problem, but operate on very different optimisation philosophies. In particular, the performance and resource use of the GA depends on the specification of the number of iterations and the number of population members. We consider different specifications of these parameters and compare the results with those of the IO method. As the GA performs many more fitness evaluations of the possible solutions compared with the IO, it is fully expected that the GA should therefore perform better than the IO, and we also expect GAs specified with higher number of population members or iterations to find the optimal solution with higher probability than GAs with lower specifications. We wanted to determine what the

variation in GA solutions for this problem would look like, and we wanted to determine how close the IO solution would get to the best available GA solution.

The coverage of monitoring sites on land regions of the Northern Hemisphere is far denser compared with the Southern Hemisphere. Previous studies on optimising locations for new monitoring sites for atmospheric CO_2 based on uncertainty reduction of surface flux estimates have identified southern Africa as an important region to constrain (Patra & Maksyutov 2002, Rayner 2004). To help meet this demand, five new instruments have been acquired by the Council for Scientific and Industrial Research in order to expand the existing network of monitoring stations in South Africa. These instruments need to be placed in optimal locations in order to maximise the investment return on the instruments through reducing the uncertainty of the estimates of CO_2 fluxes from subregions in South Africa.

In addition to the original network design problem for South Africa, which required a solution for the placement of five new stations as part of a naive observation network to solve for the aggregated CO_2 flux, we also considered two additional hypothetical scenarios. In the first scenario we considered an established network. We used the best solution for the five-member network from the available network solutions of the original problem as the starting point for the base network. We then solved for an additional five stations to add to this network. In the second scenario, we considered optimising the uncertainty reduction over only a portion of the country. We chose a region over the eastern side of South Africa which included the largest fossil fuel emitters of CO_2 and the areas of greatest biospheric activity. The purpose of these two additional scenarios was to determine how well the GA performed in relation to the IO when the observation footprints of new towers to be added to the network overlapped with those of the new and existing members of the network, or when the prior covariance matrix of the fluxes was made more complex. We hypothesized that these scenarios would lead to IO solutions that would be inferior to those from an optimally parameterised GA, and that the GA would require a greater number of evaluations in order to reach a stable solution in comparison with the original network design problem.

We present the Bayesian inverse modelling framework in the following section, and describe the optimisation algorithms and how these are used in the context of designing a measurement network aimed at reducing the uncertainty in the total flux of CO_2 from South Africa. In section 3 we present the network solutions obtained by the IO and repeated implementations of the GA algorithm at different specifications for the number of iterations and population members, followed by network solutions of the two additional scenarios. We discuss these results in section 4.

2. Methodology

2.1. Inverse Modelling Framework

The Bayesian inverse modelling approach to solve for surface fluxes was first implemented by Enting & Mansbridge (1991) based on the approach described in Tarantola (1987). This approach has since been adopted to solve for fluxes at the global scale (Bousquet et al. 1999, Gurney et al. 2003, Baker et al. 2006, Chevallier et al. 2010, Ciais et al. 2010), at the regional scale (Gerbig et al. 2003, Lauvaux et al. 2008, Broquet et al. 2013), and more recently at the city-scale (Bréon et al. 2015, Lauvaux et al. 2016). The inverse modelling framework used in this study is described in detail in Ziehn et al. (2014) and Nickless et al. (2015) and is based on the methodology outlined in Rodgers (2011), Enting (2002), and Tarantola (2005). This approach assumes that the atmospheric concentrations, $\mathbf{c}$, can be modelled based on the surface CO_2 fluxes, $\mathbf{s}$, (hereafter referred to as surface fluxes) based on the following linear relationship:

$$\mathbf{c}_{mod} = \mathbf{H}\mathbf{s} \tag{1}$$

where $\mathbf{c}_{mod}$ are the modelled concentrations at the measurement sites, and $\mathbf{H}$ is the sensitivity matrix, derived from the atmospheric transport model, which is driven by inputs derived from a regional climate model. The sensitivity matrix provides the sensitivities of modelled concentrations, $\mathbf{c}_{mod}$, to the surface fluxes ($\mathbf{s}$) (Enting 2002, Tarantola 2005). $\mathbf{H}$ maps the sources onto the observed concentrations and transforms the contribution of these sources from emission rates into CO_2 concentrations.

$\mathbf{c} - \mathbf{c}_{mod}$ represents the observation errors. These errors can be random or systematic, and can be split into measurement errors and modelling errors. Measurement errors in the CO_2 observation network, particularly those sites accredited by the Global Atmospheric Watch, have errors that are typically negligible, usually below 0.2 ppm, and can be assumed to be random. Even if $\mathbf{s}$ were perfectly known, errors in $\mathbf{c}_{mod}$ would occur due to the errors in atmospheric transport model contained in $\mathbf{H}$, whose parameters are not constrained by the inversion. Modelling errors can occur due to an imperfectly parameterised transport model, and also due to representation discrepancies, where we use a point estimate to represent the concentration of a volume, and due to aggregation, as we homogenised patchy surface fluxes within each grid cell. In reality, the air will pass over only parts of the grid cell and collect information which we relate back to the whole grid cell. Tarantola (2005) shows that modelling errors can be accounted for in the inversion by adding these to the measurement errors represented by uncertainty covariance matrix of the observations, $\mathbf{C_c}$.

If it is assumed that the observation errors and surface fluxes have unbiased Gaussian error distributions, the Bayesian cost function can be solved as follows (Enting et al. 1995, Rodgers 2011, Enting 2002, Tarantola 2005):

$$J(\mathbf{s}) = \frac{1}{2}\left((\mathbf{c}_{mod} - \mathbf{c})^T {\mathbf{C_c}}^{-1}(\mathbf{c}_{mod} - \mathbf{c}) + (\mathbf{s} - \mathbf{s_0})^T {\mathbf{C_{s_0}}}^{-1}(\mathbf{s} - \mathbf{s_0})\right) \tag{2}$$

where $\mathbf{c}$ are the observed concentrations, $\mathbf{C_c}$ is the uncertainty covariance matrix of observations, $\mathbf{s_0}$ is the vector of prior surface flux estimates, and $\mathbf{C_{s_0}}$ is the prior uncertainty covariance matrix of the surface fluxes. In regional inversion problems, the sources are usually represented by a spatial temporal grid of surface fluxes. The number of sources solved for by the inversion are then equal to the number of spatial pixels multiplied by the number of periods for which the emissions are solved. For example, if weekly fluxes are solved for by the inversion, separated into day and night sources, and if the inversion is run over a month, there will be eight periods. Therefore the number of sources solved for by the inversion is equal to the number of pixels multiplied by eight. In our South African region the domain was divided into 50×25 spatial pixels and we solved for weekly fluxes, day and night separately, therefore eight periods per month. The total number of sources which would be solved for by the inversion are 10,000. The prior and posterior uncertainty covariance matrices of these sources had dimensions 10,000×10,000.

The solution for the posterior covariance matrix of the surface fluxes, which we will use to assess the uncertainty reduction of the different network designs, is:

$$\mathbf{C_s} = \left(\mathbf{H}^T {\mathbf{C_c}}^{-1} \mathbf{H} + {\mathbf{C_{s_0}}}^{-1}\right)^{-1} \tag{3}$$

$$= \mathbf{C_{s_0}} - \mathbf{C_{s_0}} \mathbf{H}^T \left(\mathbf{H} \mathbf{C_{s_0}} \mathbf{H}^T + \mathbf{C_c}\right)^{-1} \mathbf{H} \mathbf{C_{s_0}} \tag{4}$$

The solution of the posterior uncertainty covariance matrix of the sources does not depend on the measured concentrations at the site, only on the error covariance matrix of the modelled concentrations and the prior uncertainty estimates of the surface fluxes. This makes it possible, before any concentration measurements are obtained at the site, to assess to what extent a new site can contribute to the reduction in the total uncertainty of the aggregated surface flux estimate, relative to the total uncertainty in this estimate under the base network. In our original network design problem, the base network consisted of two background measurement sites located at Cape Point and Gobabeb. $\mathbf{C_s}$ does depend on the transport matrix $\mathbf{H}$, and therefore depends on which stations are in the observation network.

The aggregation errors need to be added to the observation errors, as shown by Kaminski et al. (2001) and Tarantola (2005), and are dependent on the resolution at which the surface fluxes are solved. As shown in Nickless et al. (2015), to determine the aggregation error for each of the candidate measurement sites, the surface fluxes at a higher resolution of $0.6° \times 0.6°$ were used to represent the real, patchy surface fluxes, where these grid cells fit exactly into the grid cells of the surface solved for by the inversion. We could then use the method described in Kaminski et al. (2001) to approximate the aggregation error, where it was shown that the aggregation error $\mathbf{C}_{c,m}$ can be calculated as:

$$\mathbf{C}_{c,m} = \mathbf{H} \mathbf{P}_{-} \mathbf{C}_{s_{0h}} \mathbf{P}_{-}^T \mathbf{H}^T, \tag{5}$$

where $\mathbf{P}_{-} = \mathbf{I} - \mathbf{P}_{+}$ and $\mathbf{P}_{+}$ is the projection matrix which, if multiplied with the high

resolution surface flux estimates, $\mathbf{s}_{0h}$, produces the lower resolution flux estimates, $\mathbf{s}_0$, in positions of the corresponding high resolution fluxes. The solution of $\mathbf{C}_{c,m}$ was obtained for each measurement site, and as a conservative approach, the maximum value of the diagonal was assigned as the aggregation error for that measurement site.

We assigned values of 4 ppm^2 as the diagonal elements of $\mathbf{C_c}$. For the regional inversion performed by Wu et al. (2013), observation errors were determined to be between 2.9 and 3.6 ppm. We used the slightly lower uncertainty value of 2 ppm since our region occurs in the Southern Hemisphere where the variability in observed CO_2 concentrations is lower than in the Northern Hemisphere, and CO_2 fluxes are believed to be generally smaller as well. This value accounted for measurement error and errors in the atmospheric transport model. The aggregation errors determined using Equation 5 for each site were added to the diagonal elements. Therefore the observation errors were specific to each site.

The transport matrix, $\mathbf{H}$, is derived from a Lagrangian particle dispersion model. We have used the model developed by Uliasz (1994), which we refer to as LPDM. This model is run in backward mode for each of the potential measurement sites. A number of particles are released from the receptor point at regular intervals (every 20 seconds in this case) and tracked backwards in time. The particle counts at the surface and boundaries are recorded at each time step, where the surface is represented as a grid with a resolution of $1.2^\circ \times 1.2^\circ$ over the domain of southern Africa. The derivation of the conversion of particle counts to the influence functions which make up the elements of $\mathbf{H}$ is described in Ziehn et al. (2014) and follows Seibert & Frank (2004).

The elements of the prior uncertainty covariance matrix, $\mathbf{C_{s_0}}$, were determined by the uncertainties in the fossil fuel emissions and natural biospheric fluxes for each surface flux grid cell. The fossil fuel uncertainties were determined from ten realisations of the fossil fuel fluxes from the Fossil Fuel Data Assimilation System (FFDAS) product (Rayner et al. 2010), which is produced at a resolution of $0.1^\circ \times 0.1^\circ$. The fluxes were first aggregated to the network design resolution of $1.2^\circ \times 1.2^\circ$, and then the variances calculated for the ten realisations for each grid cell. The uncertainty values for the bioshperic fluxes, or net ecosystem productivity (NEP), were estimated as the net primary productivity (NPP) as described in Nickless et al. (2015), similar to the approach of Chevallier et al. (2010). These estimates were derived from an assessment of the natural carbon sinks of South Africa (Nickless et al. 2015). The NPP estimates were aggregated to the resolution of $1.2^\circ \times 1.2^\circ$ and these used as the uncertainty estimates of the natural fluxes. The NPP estimates are high across most of the country during the month of January, which is when most of the country has its growing season and receives the most rainfall. Conversely, during the month of July the NPP estimates are small and the majority of activity is concentrated along the Western Cape coast, which receives winter rainfall. The total flux uncertainty is dominated by biospheric flux uncertainty in January, whereas in July it is dominated by fossil fuel flux uncertainty, which is concentrated within a few small regions, mainly near major cities and power stations, spread out across the country. Figure 1 provides the spatial distribution of the

uncertainty prescribed to the NEP and fossil fuel prior fluxes.

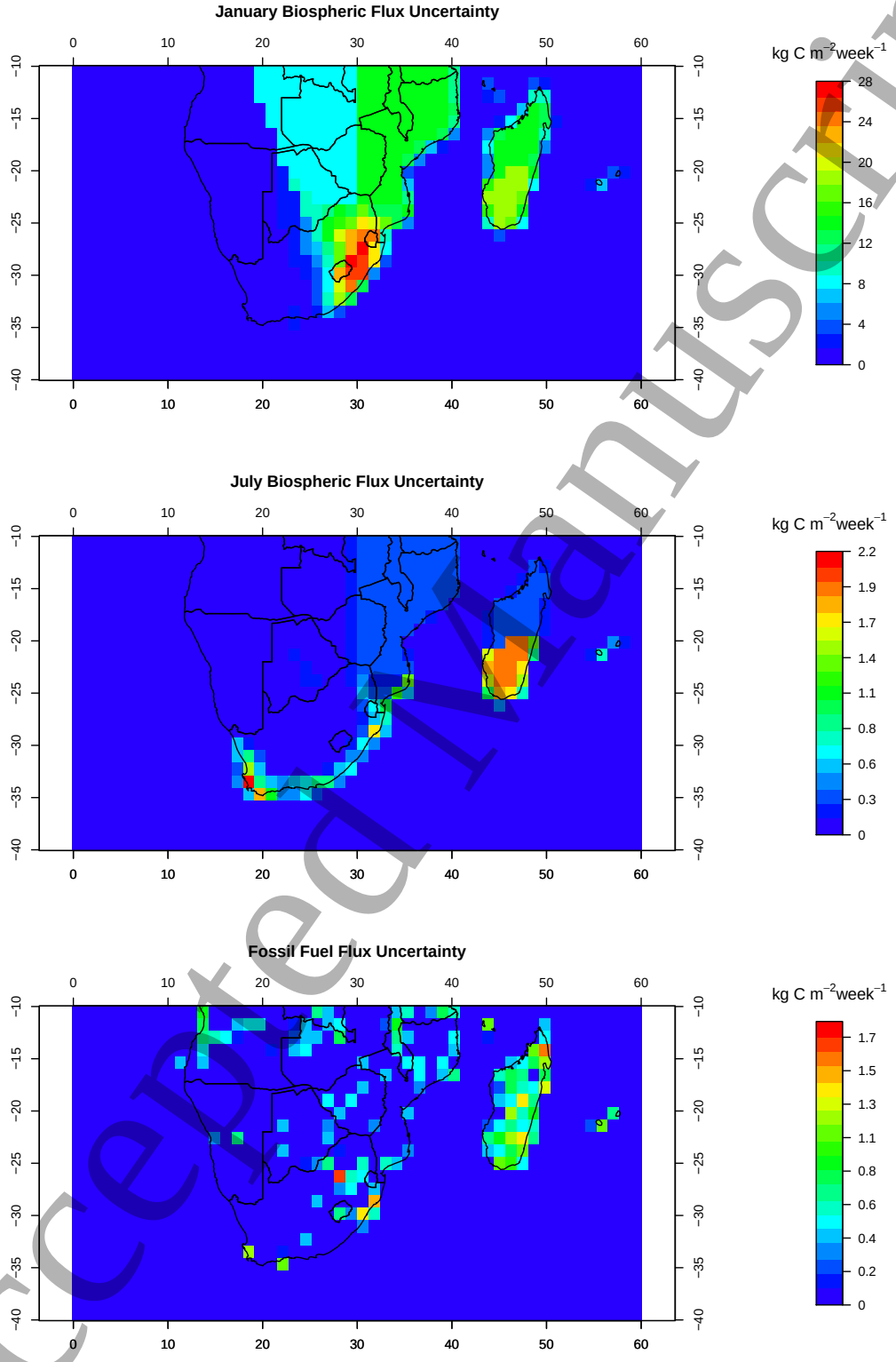

Figure 1: Prior uncertainty assigned to the net ecosystem productivity (NEP) fluxes in January and July, and the uncertainty in the fossil fuel fluxes, expressed as standard deviations (g C m^{-2} week^{-1})

Let Y represent the sum of two surface fluxes, such that $Y = s_i + s_j$, where s_i and s_j are the surface fluxes in grid cells i and j respectively. The variance of Y is equal to

$$Var(Y) = C_{s_{ii}} + 2C_{s_{ij}} + C_{s_{jj}}. \quad (6)$$

If we sum over all surface fluxes, the variance of this total is the sum of all the elements of the covariance matrix. The overall uncertainty in surface fluxes was determined by summing over all the elements of $\mathbf{C_s}$ (cost function J_{Ce}) and then taking the square root, to obtain the uncertainty of the total flux estimate for South Africa expressed as a standard deviation:

$$J_{Ce} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} C_{s_{ij}}} \quad (7)$$

The cost function used to assess the network members was based on the uncertainty reduction calculated as:

$$U_R = 1 - \frac{\hat{J}_{Ce}}{J_{Ce\,base}} \quad (8)$$

where $\hat{J}_{Ce}$ is the posterior standard deviation of the total flux estimate of the proposed network containing the additional five stations, and $J_{Ce\,base}$ the standard deviation estimate of the total estimate of the base network, which is determined from the posterior covariance matrix of the surface fluxes if only the existing Cape Point and Gobabeb stations are in the network. The use of the uncertainty reduction as a proportion of the prior uncertainty was used by Rayner (2004). Here the trace of $\mathbf{C_s}$ was used instead of the sum of the covariance elements to represent the total uncertainty. We performed a sensitivity analysis that compared these approaches, and found that the resulting network solutions were similar. A more generalised approach to optimal experimental design in inverse problems in provided in Haber et al. (2008), where they use the trace of the posterior covariance matrix, and show how this can be used to optimise measurement networks for inverse problems using alternative methods of regularisation.

The candidate stations and the existing background stations are displayed in Figure 2.

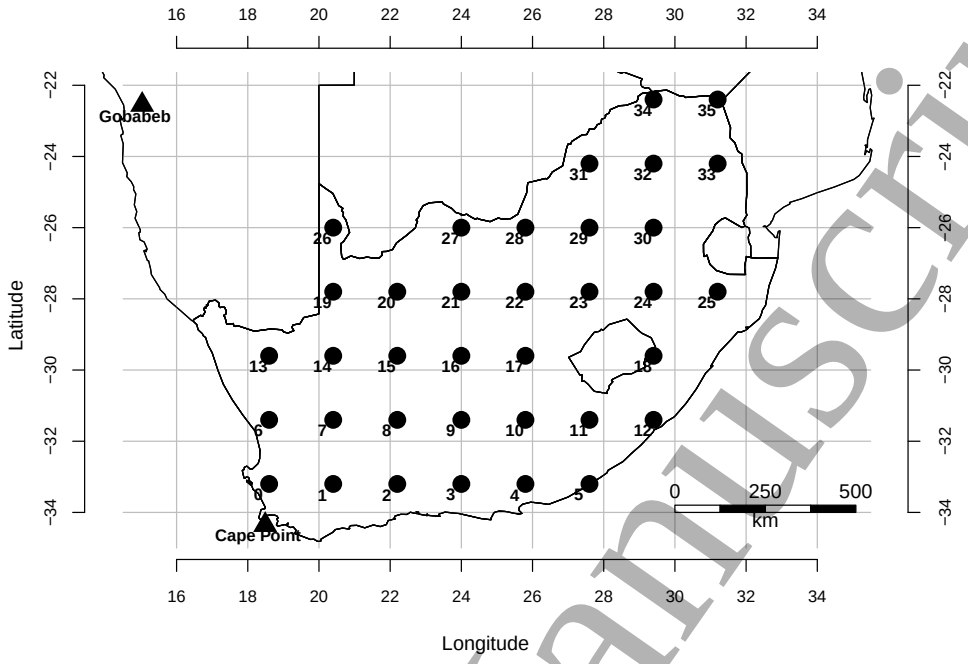

Figure 2: Location of candidate stations within South Africa together with the existing background stations at Cape Point and Gobabeb. Of these candidate stations, five locations are required for the placement of new monitoring sites for atmospheric CO_2 concentrations to solve the original network design problem.

2.2. Optimisation Routines

2.2.1. Incremental Optimisation A description of the use of Incremental Optimisation (IO) to solve general optimisation problems is provided by Hartline & Sharp (2006). IO for CO_2 monitoring network design problems was proposed by Patra & Maksyutov (2002), where the addition of new stations to an existing measurement network was considered. The optimal network was thought of as a problem which consisted of several subproblems, where each was solved incrementally. Given an existing network, the first subproblem was to determine the first station to add to the network that would result in the greatest reduction in uncertainty. The candidate list of stations was derived from equally spaced locations within South Africa; a total of 36 stations (Figure 2). Adding each of the candidate stations in turn to the base network, the cost function was calculated, and the candidate station yielding the greatest uncertainty reduction was removed from the candidate list and added to the base network list. This procedure was repeated until the desired network size was reached. In this case the network size would be seven - two existing and five new stations.

The IO routine results in an evolution of the network solution, allowing the user to determine the best station to add to the network at each iteration, as well as the uncertainty reduction that each of the unselected stations would have had, providing a list of best alternatives. In addition, the IO routine is computationally inexpensive. This is an advantage due to the already high computational costs of solving for the posterior

covariance matrix that would need to be calculated for each potential solution.

2.2.2. Genetic Algorithm The genetic algorithm (GA) draws on the concept of 'survival of the fittest' from evolutionary biology to determine the best solution for a numerical optimisation problem, and has a wide range of applications (Chambers 2001). GA optimisation has been used to solve various network design problems, such as placement of wind turbines to maximise the power output (Grady et al. 2005) and the design of a telecommunications network to support the expected communications traffic between pairs of connections (Berry et al. 1999). For our optimisation problem the solution would be a list of stations to add to the existing network to best reduce the surface flux uncertainty. Each of the five stations in the network solution represents a parameter in the algorithm. The GA procedure carried out in this study is based on Rayner (2004).

The GA does not evolve a single solution, but rather a population of solutions. The population is made up of a number of potential five station solutions for the network problem. Each solution represents a member of the population. Through a process of culling, mutation and reproduction, population members are lost, evolved and replicated for the next iteration, all based on pseudo-random numbers generated from the uniform distribution. The algorithm begins by randomly generating a population of solution members of size N, where each solution consists of a list of five stations, based on the candidate list of stations; the same candidate list used by the IO. For this network design, candidate stations may appear only once in a solution. The first part of a GA iteration is the pairwise swap over of parameter values (i.e. stations) between two randomly selected population members, determined by whether a random uniform number is between zero and a pre-selected cross-over probability. Individual parameter values are also changed based on a random uniform number and the specified mutation probability. Based on a criterion for fitness, calculated as:

$$F = 1 - \frac{r - 0.5}{N} \tag{9}$$

where r is the ordinal ranking of the member according to each member's cost function and N is the population size, solutions are removed (or culled) from the population of solutions if a pseudo-random number generated from the uniform distribution for each of the population members is greater than F. The resulting probability of a member being culled is therefore 1 - F.

Once the culling procedure is completed, new members are added to the population through a process of reproduction until the population size is back to N. The current population members are repeated if a randomly generated number is below the member's fitness score. The probability of a member being replicated in the new generation is therefore equal to F. Sampling of the members is with replacement, so those population members with good fitness scores will have a better chance of appearing multiple times in the population of solutions. To ensure that the diversity of the solutions in the population is high enough to avoid the algorithm getting stuck at a local extrema, an additional process of mutation is applied to the population members. Here a random

uniform number is generated for every parameter of each population member. If the value is below the mutation threshold, then the current value of the parameter is replaced from a randomly selected parameter from the candidate list of stations. We used the recommended mutation rate of 0.1 suggested by Rayner (2004). This concludes one iteration of the GA. The algorithm iterates until the pre-determined number of iterations is achieved.

Elitism is maintained, so that set of five candidate stations with the highest fitness is replicated unchanged into the next generation of population members at every iteration. Once the algorithm is complete, the member of the population with the best score is selected as the final solution.

2.3. Optimal Network Comparisons

For two representative months, January for summer and July for winter, the transport matrix, $\mathbf{H}$, was derived from the particle counts generated by LPDM, and the prior covariance matrices for weekly surface fluxes, $\mathbf{C_{s_0}}$, for these periods were constructed. The IO routine was run to determine the optimal network for the two months separately. The optimisation was repeated, using the GA under different specifications of the number of iterations and the population size. Since the outcome of the algorithm has the potential to be inconsistent from run to run, due to the use of pseudo-random numbers, the algorithm was run five times for each configuration. The purpose of this was to demonstrate some of the variability in the GA solution set. The configurations considered were population sizes of 50 or 100 with either 50, 75 or 100 iterations. These configurations therefore result in GAs with different numbers of evaluations. For example, a GA with 50 population members and 50 iterations would have 2,500 evaluations whereas a GA with 100 population members and 100 iterations would have 10,000 evaluations. We expected the solution set from the GA to stabilise as the number of evaluations increased. The time taken to run each of the algorithms was recorded in order to compare the efficiency of the algorithms.

The utility (or fitness) of a network of monitoring stations was assessed by means of the percentage uncertainty reduction in the total posterior flux estimate for the region relative to the uncertainty of the base network, as calculated from Equation 8. We expected the IO to obtained an inferior solution to the GA, and we wanted to assess how the IO solution compared in fitness to the best available GA solution. How similar two networks were in terms for their placement was assessed using the dissimilarity index (DI). How similar two network solutions were was of interest, because if two solutions had identical uncertainty reductions, but very different placement of stations, considerations such as the cost of establishing the network and the feasibility of the network in practice would need to be used to make a decision on which network would be implemented. The DI was calculated as the sum of the distance to the nearest neighbour in the compared

network, over all the members in the pair of assessed networks.

$$DI = \sum_{i=1}^{5} \min_{j} \sqrt{\Delta x_{ij}^2 + \Delta y_{ij}^2} + \sum_{j=1}^{5} \min_{i} \sqrt{\Delta x_{ij}^2 + \Delta y_{ij}^2} \quad (10)$$

where i and j $\in$ [1,2,3,4,5], and Δx_{ij}^2 and Δy_{ij}^2 are the squared differences between the Cartesian coordinates of the i^{th} station in the first network and the j^{th} station in the second network. The first term in the DI calculation is the sum of the nearest neighbour distances between each site in the first network solution and its nearest neighbour in the second solution set (a sum of five nearest neighbour distances). The second term sums the nearest neighbour distances between each station in the second network solution and its nearest neighbour in the first network solution (a sum of five nearest neighbour distances). Therefore the DI is the sum of ten nearest neighbour distances when comparing two five-member networks.

In cases where the two networks compared were the same, the index results in a value of zero. Networks which did not contain exactly the same members would always have a DI of greater than zero. The index increases as the networks become more dissimilar in space. This provides a one-number measure of network similarity that can consistently be used for the network comparisons, provided each solution consists of the same number of stations, and allows for an objective assessment of how different the positioning of sites are between two network solutions which may not be obvious to the eye. The index provides a measure of distance between solutions in terms of kilometres. The distribution characteristics of the DI if it were used to compare two randomly selected five-member networks are provided in the appendix.

To assess the degree to which sites in a particular network solution were clumped together, we calculated the centroid of the spatial points from each network solution, and calculated the mean distance between each station in the solution and this centroid. The centroid provides additional information on where in South Africa the stations were concentrated. We measured the degree of clumping or clustering of stations in order to assess if the network solutions were concentrated over a particular region, or if the stations in the network solution were spread over the domain.

2.4. Additional Scenarios

Optimisations were performed for the original network problem, which required five additional stations to add to the existing background stations at Cape Point and Gobabeb. In this scenario it would be expected that the optimal solution would be one where the observation footprints of the towers are not overlapping, in order to view us much of the uncertainty across the country as possible. Both GA and IO should be able to achieve solutions which we would not expect to differ greatly, but as the optimisation problem was more complex than the original problem, the GA could potentially find solutions that were inaccessible through the IO.

We considered two additional hypothetical scenarios. In the first of these scenarios we considered the situation which may be in existence once the five new stations are

established. The base network under this scenario had seven stations (two background stations and five new stations), and the purpose of the network design was to find a solution for the placement of five additional stations, to take the network to 12 stations. We used the best solution out of all available solutions from the original network problem for these five existing stations, and the optimisation routine was used to solve for the five-member network to add to this established network. Under these conditions, the new stations would more likely have observation footprints which were overlapping with those from the new elements and old elements of the network. The observation footprint of a site would depend on the prevailing atmospheric conditions around that site, described by the sensitivity matrix. Compared to the original network problem, which aimed to reduce the uncertainty of the total flux estimate that was almost entirely unconstrained by the base network, the established network design problem had the objective of applying additional constraint to a network that already covered the main sources of uncertainty. Therefore, the aim of the established network design problem was to plug the holes in an existing network. The prior covariance matrix of the fluxes would be more complex under this scenario and we wanted to determine if the GA, which considers the global parameter solution, would be better suited than the IO to find an optimal solution under these circumstances, where assessing the additional stations simultaneously rather than sequentially could be an advantage.

The second hypothetical scenario was to adjust the original network problem so that the uncertainty reduction resulting from the network was optimised over a subregion of the country, rather than over the total land surface of South Africa. This was achieved by summing the variance and covariance terms of the posterior covariance matrix, $\mathbf{C_s}$, for only those elements related to the subregion. The region we selected comprises a large grid on the eastern half of South Africa which includes the largest fossil fuel emitters and the largest biospheric sinks of CO_2. Under this scenario, the observation footprints of the measurement towers would be closer and may overlap in order to view the uncertainty over this subregion. The optimal network solution should be dependent on how much overlap (i.e. redundancy) can be afforded. A solution may require two sites close to each other in order to get a more comprehensive view of a source with large uncertainty. There would then be a trade-off between reducing the large uncertainty from this source and the lost opportunity to view elsewhere in the domain not already covered by the network. Stations did not necessarily need to be located within the subregion to reduce the subregion uncertainty. We wanted to assess if the GA would be better suited to optimising the overlap between observation footprints compared with the IO method to obtain a better uncertainty reduction in the overall flux in the subregion.

2.5. System configuration

The optimisation routines were carried out using Python version 2.7 on a desktop computer with a Linux operating system and a Intel Core i7-3770K processor running

at 3.50 GHz with four cores and eight logical processors, and 16 gigabytes of RAM.

3. Results

3.1. Original Network Design Problem

Table 1 gives the uncertainty reduction results of the optimal network solutions under the different algorithm runs, as well as the mean distances to the centroid, DI's with the IO solution, and run times. The uncertainty reduction ranged between 76.5% and 78.8% in January, and between 42.9% and 43.3% in July. The IO, for both January and July, did not find the global maximum, but a local maximum with uncertainty reduction very close to the best GA solution, differing by an absolute amount of only 0.5% in January and 0.3% in July. As these are percentage uncertainty reductions, given all of the assumptions needed for the inversion solution, networks differing by 1%, or even up to 5%, in uncertainty reduction may be considered to have the same utility. The uncertainty reductions obtained by the different GA network solutions were very similar, a difference of only 0.4%. As expected, the consistency in the GA solutions increased as the number of evaluations increased. Comparing between months, the consistency of the GA solutions in July was much higher compared with the GA solutions in January.

The mean distance to the centroid will reduce as all sites are located nearer and nearer to a central point. As more sites are located far from the central point, this metric will increase. The mean distances to the centroid for the IO solution when compared with the best GA result were not the same for either January or July, indicating that the clumping of stations differed between the IO and GA solutions, with the best GA solutions having slightly higher values for the mean distance to the centroid indicating more spread in the placement of sites. The mean distance to the centroid was generally greater for January solutions compared with July solutions, indicating that stations were more spread out for the network solutions obtained for the winter month than for the summer month. We would expect this as the network would need to have stations concentrated on the eastern side of South Africa to cover both the high uncertainty in the biospheric fluxes occurring here during summer, and the large contribution of fossil fuel emissions. In the winter months the uncertainty would be mainly due to fossil fuel emissions, and therefore the network would need to view the cities such as Johannesburg, Pretoria and Durban, and also Cape Town to the south west of the country. Therefore sites would be spread out in order to reduce the fossil fuel emission uncertainty.

In January and July, the dissimilarity index for IO was greater than zero when compared with the best GA result, with a value of 772 km in January and 879 km in July. For the current set of candidate sites, a comparison between two randomly generated networks would on average result in a dissimilarity index of 2763 km, with a standard deviation of 720 km, determined through simulation methods. A histogram showing the distribution of the mean dissimilarity index for two random networks is given in the appendix. The lower limit of the normal range (approximately the 2.5^{th}

percentile) for the DI of randomly generated networks is 1350 km. The DI's obtained between the GA solutions and the IO solution were all below this value, which suggests that IO and GA solutions were more similar than two randomly selected solutions (Table 1). This provides evidence that both optimisation routines were aiming towards similar solutions.

The time taken to obtain a result ranged between 5 and 18 times longer under the GA compared with under the IO, with the largest run times associated with a population size of 100 and 100 iterations. Increasing the population size for the GA from 50 to 100 or increasing the iterations from 50 to 100 led to roughly doubling of the run time. Increasing both the number of iterations and population size to 100 resulted in a run time between 3.6 and 3.8 times longer than the initial GA setup. Therefore the computational cost of even the lowest configured GA is substantially larger compared with the IO routine.

The GA run under the initial setup of 50 iterations and 50 population members produced consistent network solutions for July but not for January, where all five of the July GA solutions were the same but only two of the five January solutions were the same. Greater inconsistencies were observed in solutions for January across the different GA configurations. In January the spatial distribution of fluxes, particularly biospheric fluxes, was far more dispersed across the whole of South Africa, and the prior uncertainties in the surface fluxes were much larger in January compared with July. This result is illustrated using the heat maps presented in Figures 3a and 3b. Much larger variability in the DI's is apparent when comparing GA solutions under fewer population members and iterations compared with GA solutions under larger population sizes or with greater number of iterations, as indicated by the lighter blocks in the heat map for January in the top right hand corner. This implies that for the month of January, the GA was converging towards a solution under larger population sizes and number of iterations. This convergence occurred more readily for July, where the heat map already showed lighter blocks compared with the initial setup just from increasing the iterations from 50 to 75. Across the different configurations of the GA the DI values for July were always lower compared with those for January (Figures 3a and 3b).

The best solution was not obtained under the cheapest GA configuration for either January or July. For the month of January when the number of iterations was increased to 75, two of the five GA runs converged to this solution, but when increased to 100 iterations, three of the five solutions converged to this solution. As expected, increasing the population size to 100 improved the chances of obtaining this best solution relative to the initial configuration, and this probability also increased as the number of iterations increased. But only two of the five solutions obtained the best solution when the population size and number of iterations were 100. Under this configuration the DI values between the network solutions were the lowest relative to comparisons within any of the other GA configurations (Figure 2a) Therefore, even when these two GA parameters were set relatively high, there was no guarantee of obtaining the best solution under the January scenario, although the differences between network solutions

did decrease. In January the surface flux uncertainty was high, relative to July, and dispersed unevenly across South Africa, particularly concentrated in the northern and eastern parts of the country where the greatest biological activity takes place in Summer, but also where the large cities are located. Figures 4a and 4b provide the locations of all the sites from the different network solutions, together with the NPP surface fluxes (representative of the biospheric flux uncertainty) for January and July respectively. In January the sites of the different solutions are concentrated towards the central and eastern parts of the country, supporting the lower mean distances to the centroid obtained in January. Some sites are consistently selected, such as site 27 located at the border of South Africa and Zimbabwe, or site 11 located south of Lesotho. Other sites appear far less frequently, such as site 28 and site 10. Both these sites are located one step away from the most frequently selected sites, and therefore would be in a position to view much of the same uncertainty. In general, the network solutions have placed sites next to locations with the highest uncertainty in either fossil fuel or biogenic surface fluxes.

In July, the best GA solution was obtained by increasing either the number of iterations or the population size. By increasing the number of iterations to 75, four out of the five network solutions resulted in the best GA solution. Under the different configurations with iterations or population size greater than 50, between two and four of the network solutions converged to the best solution, with the maximum configuration resulting in four out of five network solutions as the best solution. Under the July scenario, the overall uncertainty in the surface fluxes was much lower compared with January, and the largest uncertainties were concentrated towards the south west, where the greatest amount of biological activity would be taking place in winter and where the city of Cape Town is located, but also around the cities of Johannesburg and Pretoria in the central part of the country, and Durban to the south east. Figure 4b illustrates the network solutions for July. There was far less variability in the placement of stations compared with January. The stations were now split between those that view uncertainty around Cape Town and in the northern parts of the Western Cape province, where fossil fuel emission uncertainty and biogenic flux uncertainty are both high, and between those on the eastern side of the country, located near fossil fuel sources.

Table 1: Table of uncertainty reductions achieved by each network in the original network design problem solving for five station locations, together with the network's centroid point, the mean distance to centroid, dissimilarity index, and running times. Results are presented for the IO result, and results under the GA with 50, 75, and 100 iterations together with 50 population members (GA 50 50, GA 75 50, GA 100 50), and with 50, 75 and 100 iterations together with 100 population members (GA 50 100, GA 75 100, GA 100 100)

	January					July				
Algorithm Specification	Uncertainty kt C week^{-1}	Uncertainty reduction (%) (mean)	Centroid, mean distance to centroid (km)	DI with IO (km)	Run time (min)	Uncertainty kt C week^{-1}	Uncertainty reduction (%) (mean)	Centroid, mean distance to centroid (km)	DI with IO (km)	Run time (min)
Prior	2058.99					102.79				
IO	447.21	78.28	(27.6, -28.88), 303	0	14	58.66	42.933	(24, -29.96), 502	0	12
GA 50 50	438.09	78.723	(27.6, -28.52), 310	1115	68	58.66	42.933	(24, -29.96), 502	0	75
GA 50 50	440.25	78.618	(27.6, -28.88), 364	1132	68	58.66	42.933	(24, -29.96), 502	0	74
GA 50 50	459.35	77.69	(26.88, -28.88), 363	1206	65	58.66	42.933	(24, -29.96), 502	0	74
GA 50 50	438.09	78.723	(27.6, -28.52), 310	1115	73	58.66	42.933	(24, -29.96), 502	0	78
GA 50 50	442.42	78.513 (78.453)	(27.24, -29.24), 334	371	68	58.66	42.933 (42.933)	(24, -29.96), 502	0	78
GA 75 50	438.09	78.723	(27.6, -28.52), 310	1115	98	58.29	43.294	(24.72, -29.6), 507	879	109
GA 75 50	483.83	76.502	(27.24, -28.52), 352	1542	97	58.29	43.294	(24.72, -29.6), 507	879	109
GA 75 50	436.23	78.813	(27.24, -28.88), 357	772	105	58.29	43.294	(24.72, -29.6), 507	879	110
GA 75 50	438.66	78.696	(27.24, -28.88), 326	1226	100	58.29	43.294	(24.72, -29.6), 507	879	109
GA 75 50	436.23	78.813 (78.309)	(27.24, -28.88), 357	772	99	58.66	42.933 (43.222)	(24, -29.96), 502	0	110
GA 100 50	436.23	78.813	(27.24, -28.88), 357	772	130	58.29	43.294	(24.72, -29.6), 507	879	159
GA 100 50	436.23	78.813	(27.24, -28.88), 357	772	128	58.29	43.294	(24.72, -29.6), 507	879	144
GA 100 50	438.09	78.723	(27.6, -28.52), 310	1115	133	58.66	42.933	(24, -29.96), 502	0	144
GA 100 50	436.23	78.813	(27.24, -28.88), 357	772	130	58.29	43.294	(24.72, -29.6), 507	879	144
GA 100 50	438.09	78.723 (78.777)	(27.6, -28.52), 310	1115	133	58.66	42.933 (43.15)	(24, -29.96), 502	0	148
GA 50 100	436.24	78.813	(27.6, -28.16), 363	802	138	58.66	42.933	(24, -29.96), 502	0	148
GA 50 100	438.09	78.723	(27.6, -28.52), 310	1115	132	58.29	43.294	(24.72, -29.6), 507	879	147
GA 50 100	436.23	78.813	(27.24, -28.88), 357	772	105	58.29	43.294	(24.72, -29.6), 507	879	146
GA 50 100	438.66	78.696	(27.24, -28.88), 326	1226	100	58.29	43.294	(24.72, -29.6), 507	879	155
GA 50 100	438.66	78.696 (78.748)	(27.24, -28.88), 326	1226	99	58.29	42.933 (43.15)	(24.72, -29.6), 507	879	170
GA 75 100	436.23	78.813	(27.24, -28.88), 357	772	189	58.66	42.933	(24, -29.96), 502	0	223
GA 75 100	437.06	78.773	(27.24, -28.52), 396	1173	194	58.29	43.294	(24.72, -29.6), 507	879	215
GA 75 100	436.24	78.813	(27.6, -28.16), 363	802	196	58.29	43.294	(24.72, -29.6), 507	879	214
GA 75 100	436.23	78.813	(27.24, -28.88), 357	772	210	58.29	43.294	(24.72, -29.6), 507	879	218
GA 75 100	438.09	78.723 (78.787)	(27.6, -28.52), 310	1115	209	58.29	43.294 (43.222)	(24.72, -29.6), 507	879	233
GA 100 100	436.24	78.813	(27.6, -28.16), 363	802	252	58.66	42.933	(24, -29.96), 502	0	280
GA 100 100	436.23	78.813	(27.24, -28.88), 357	772	254	58.29	43.294	(24.72, -29.6), 507	879	284
GA 100 100	436.24	78.813	(27.6, -28.16), 363	802	257	58.29	43.294	(24.72, -29.6), 507	879	282
GA 100 100	436.24	78.813	(27.6, -28.16), 363	802	254	58.29	43.294	(24.72, -29.6), 507	879	286
GA 100 100	436.23	78.813 (78.813)	(27.24, -28.88), 357	772	256	58.29	43.294 (43.222)	(24.72, -29.6), 507	879	287

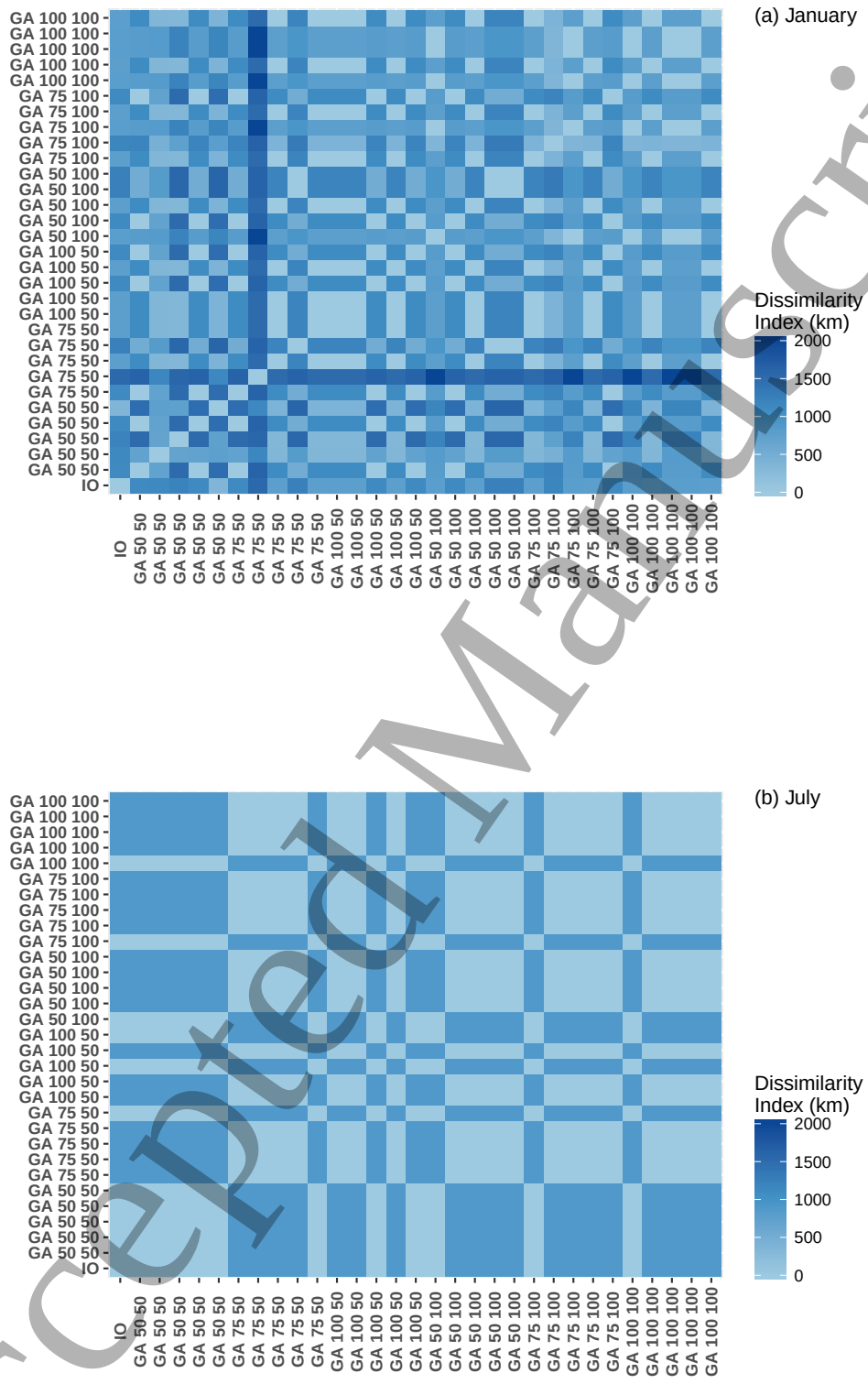

Figure 3: Heat maps displaying Dissimilarity Indices (DIs) for the January (top) and July (bottom) optimal network solutions. Network solutions which have similar placement of stations in space will have DIs closer to zero. The abbreviated names for the GA runs are the same as for Table 1.

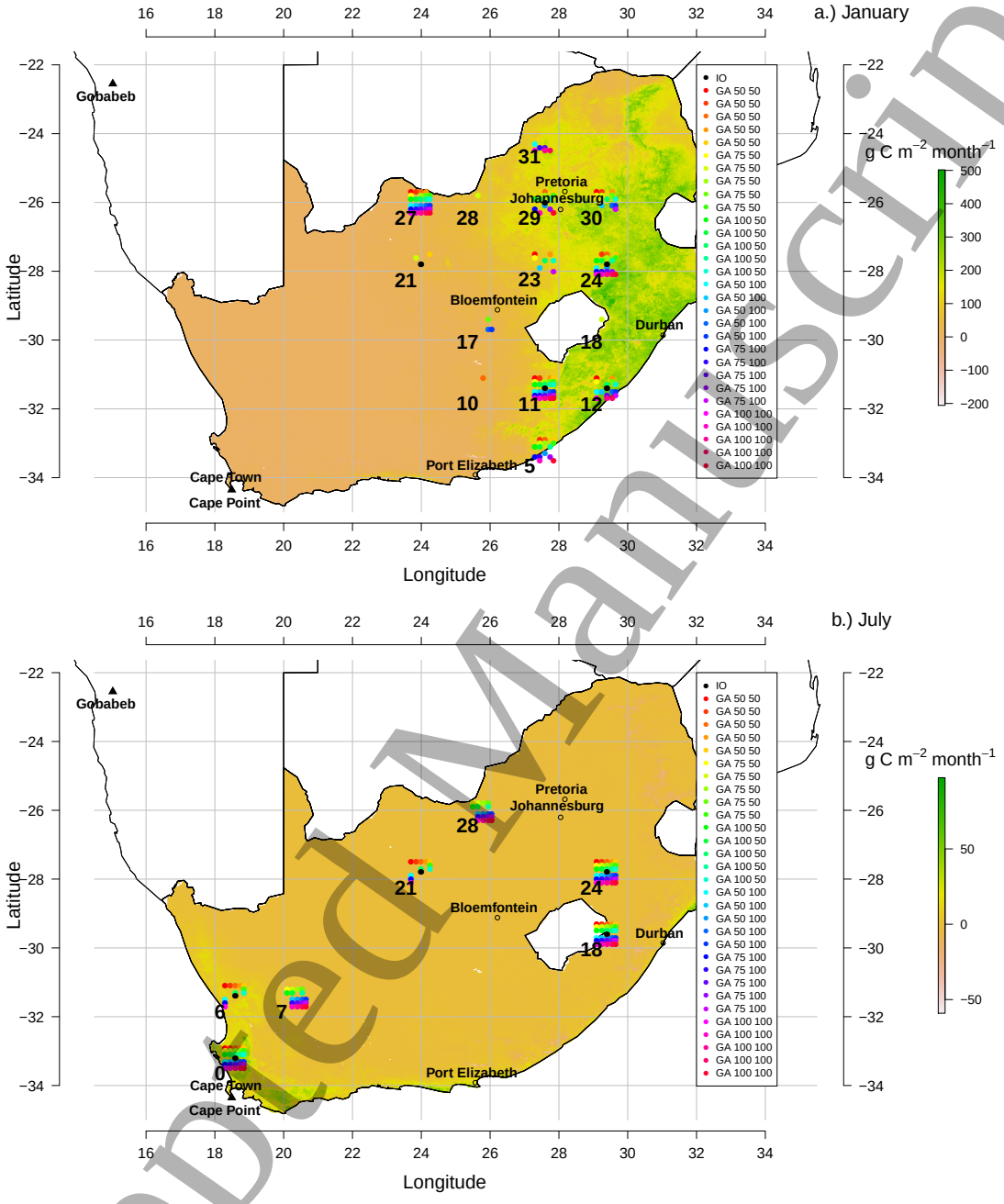

Figure 4: Maps of the optimal network locations for (a) January and (b) July from each of the algorithm runs, overlaid on the net primary productivity (NPP) (g C m^{-2} month^{-1}). Code numbers for the GA solutions are the same as for Table 1. Numbers appearing on the maps are the station locations which have appeared in one or more network solutions. Black triangles - existing network stations of Cape Point and Gobabeb. Open circles - Major South African. Black closed circles - IO network solution. Coloured closed circles - GA solutions. Each colour represents a different GA solution. Points are laid out row by row, with the top row corresponding to the GA 50 50 and the bottom corresponding to the GA 100 100.

3.2. Established Network Design Problem

In the established network scenario, the percentage uncertainty reduction achieved by each network solution was lower compared with the original network problem. In January the uncertainty reduction was approximately 78% for the original network problem, whereas for the established network solution, the reduction in the uncertainty remaining after a five-member network was established could only be improved by 50% (Table 2). Similarly in July, the original network solution was able to achieve an uncertainty reduction in the region of 43%, but once a network was established the remaining uncertainty in the total flux solution could only be reduced by 14%. This shows the diminishing returns of adding more stations to the network.

For the month of January the GA procedure was able to achieve a better solution by 3% compared to the solution obtained by IO. The GA configured with 50 iterations and 50 population members obtained the best solution as well as the worst solution with an uncertainty reduction of 45.2%. The IO solution reduced the uncertainty by 46.9%, whereas the best solution GA was able to obtain an uncertainty reduction of 49.9% (Table 2). This solution was obtained more reliably when the number of iterations and population size were at higher settings, with all GA configurations with 100 iterations obtaining the best solution. In July the GA always found a better solution compared with IO. The IO obtained a solution with 13.7% uncertainty reduction, whereas the solutions from the GA showed very little variation and all solutions with 14.2% uncertainty reduction. Figure 5 shows the cumulative influence of the surface sources on the concentration observations observed at the sites in the best solution for each month and for each of the scenarios. These maps show which sources (i.e. which of the surface pixels) are in view of the measurement sites in the best solution. As expected, the observation footprints of the measurement towers for the solutions under the established network design problem showed some overlap; more so than under the original network design problem (Figure 5). Under the established network design solutions, there is greater coverage of the coastal cities, such as Port Elizabeth and Durban.

The mean distances to the centroid and the map of the network solution show that more clustering of the stations towards a central point occurred in the January solutions than the July solution, with mean distances to the centroid of between 295 and 338 km in January verses 526 km on July, but the clustering in January was to a lesser extent compared with that of the original network problem which was between 303 and 357 km (Table 2 and Figure 6). The map of the January solutions indicates that stations tended to be placed along the central vertical line of the country, between the area of lowest uncertainty and the area of highest uncertainty, with one station located near the largest cities (Johannesburg and Pretoria) and power generation sites (Figure 6). The July solutions also tended to place stations along a boundary, this time between areas of high winter biogenic activity in the Western Cape and low activity in the rest of the country, with two stations located near the cites of Johannesburg and Pretoria;

one north and one south of these two cities (Figure 6).

The pattern of DI values revealed that there was much more consistency between the solutions from the GA compared with the original network design problem, with more dissimilarity occurring in January, when biospheric flux uncertainty was higher and more spread out across the eastern side of the country (Figure 7). The DI values for the July solutions show almost no variation between the network solutions, with GA solutions for iterations at 75 or above, or population size at 100, all resulting in the same network solution.

Run times to complete the optimisation routines were slightly higher compared with the original network problem, as the sensitivity matrix ($\mathbf{H}$) and observation error covariance matrix ($\mathbf{C_c}$) for the established scenario were larger in size due to a larger number of monitoring stations, and therefore larger number of hourly observations. The relative time differences to complete an optimisation between the IO and the different GA configurations remains similar to the original network problem.

In both January and July, under the established network scenario, the GA was able to find a better solution compared with the IO method, but the IO was within 3% of the uncertainty reduction achieved by any of the GA solutions. Only the GA at the lowest configuration had one solution with lower uncertainty reduction compared with the IO for the month of January. The improvement of the GA best solution over the IO was small; more so in July where the regions of high uncertainty were concentrated over small areas spread out across the domain. The spatial differences in these solutions were small, as evidenced by the small DI values.

Table 2: Table of uncertainty reductions achieved by each network in the extended network design problem, together with the network's centroid point, the mean distance to centroid, dissimilarity index, and running times. Results are presented for the IO result, and results under the GA with 50, 75, and 100 iterations together with 50 population members (GA 50 50, GA 75 50, GA 100 50), and with 50, 75 and 100 iterations together with 100 population members (GA 50 100, GA 75 100, GA 100 100), where solutions were generated under the assumption that a seven-member network already existed, which included Cape Point and Gobabeb and the best five-member solution from the original network design problem.

			January			July				
Algorithm Specification	Uncertainty kt C week^{-1}	Uncertainty reduction (%) (mean)	Centroid, mean distance to centroid (km)	DI with IO (km)	Run time (min)	Uncertainty kt C week^{-1}	Uncertainty reduction (%) (mean)	Centroid, mean distance to centroid (km)	DI with IO (km)	Run time (min)
Prior	436.23					58.29				
IO	231.42	46.949	(26.88, -29.24), 315	0	15	50.3	13.701	(23.28, -30.32), 427	0	14
GA 50 50	219.61	49.657	(26.52, -28.52), 305	941	96	49.99	14.238	(24.72, -29.96), 526	839	98
GA 50 50	239	45.212	(26.88, -29.24), 295	802	93	50.02	14.178	(24.72, -29.24), 490	775	99
GA 50 50	218.41	49.932	(26.52, -28.88), 338	541	91	49.99	14.238	(24.72, -29.96), 526	839	98
GA 50 50	218.41	49.932	(26.52, -28.88), 338	541	91	49.99	14.238	(24.72, -29.96), 526	839	97
GA 50 50	218.41	49.932 (48.933)	(26.52, -28.88), 338	541	93	49.99	14.238 (14.226)	(24.72, -29.96), 526	839	98
GA 75 50	228.37	47.649	(26.52, -29.24), 352	941	133	49.99	14.238	(24.72, -29.96), 526	839	138
GA 75 50	218.41	49.932	(26.52, -28.88), 338	541	133	49.99	14.238	(24.72, -29.96), 526	839	133
GA 75 50	227.95	47.745	(26.52, -28.88), 319	1342	132	49.99	14.238	(24.72, -29.96), 526	839	132
GA 75 50	218.41	49.932	(26.52, -28.88), 338	541	133	49.99	14.238	(24.72, -29.96), 526	839	135
GA 75 50	218.41	49.932 (49.038)	(26.52, -28.88), 338	541	133	49.99	14.238 (14.238)	(24.72, -29.96), 526	839	133
GA 100 50	218.41	49.932	(26.52, -28.88), 338	541	176	49.99	14.238	(24.72, -29.96), 526	839	195
GA 100 50	218.41	49.932	(26.52, -28.88), 338	541	178	49.99	14.238	(24.72, -29.96), 526	839	195
GA 100 50	218.41	49.932	(26.52, -28.88), 338	541	175	49.99	14.238	(24.72, -29.96), 526	839	196
GA 100 50	218.41	49.932	(26.52, -28.88), 338	541	177	49.99	14.238	(24.72, -29.96), 526	839	183
GA 100 50	218.41	49.932 (49.932)	(26.52, -28.88), 338	541	176	49.99	14.238 (14.238)	(24.72, -29.96), 526	839	185
GA 50 100	218.41	49.932	(26.52, -28.88), 338	541	177	49.99	14.238	(24.72, -29.96), 526	839	173
GA 50 100	219.61	49.657	(26.52, -28.52), 305	941	174	49.99	14.238	(24.72, -29.96), 526	839	171
GA 50 100	218.41	49.932	(26.52, -28.88), 338	541	179	49.99	14.238	(24.72, -29.96), 526	839	173
GA 50 100	218.41	49.932	(26.52, -28.88), 338	541	176	49.99	14.238	(24.72, -29.96), 526	839	174
GA 50 100	218.41	49.932 (49.877)	(26.52, -28.88), 338	541	175	49.99	14.238 (14.238)	(24.72, -29.96), 526	839	176
GA 75 100	218.41	49.932	(26.52, -28.88), 338	541	260	49.99	14.238	(24.72, -29.96), 526	839	257
GA 75 100	218.41	49.932	(26.52, -28.88), 338	541	260	49.99	14.238	(24.72, -29.96), 526	839	255
GA 75 100	218.41	49.932	(26.52, -28.88), 338	541	260	49.99	14.238	(24.72, -29.96), 526	839	258
GA 75 100	218.41	49.932	(26.52, -28.88), 338	541	263	49.99	14.238	(24.72, -29.96), 526	839	256
GA 75 100	218.41	49.932 (49.932)	(26.52, -28.88), 338	541	260	49.99	14.238 (14.238)	(24.72, -29.96), 526	839	253
GA 100 100	218.41	49.932	(26.52, -28.88), 338	541	342	49.99	14.238	(24.72, -29.96), 526	839	339
GA 100 100	218.41	49.932	(26.52, -28.88), 338	541	345	49.99	14.238	(24.72, -29.96), 526	839	337
GA 100 100	218.41	49.932	(26.52, -28.88), 338	541	341	49.99	14.238	(24.72, -29.96), 526	839	340
GA 100 100	218.41	49.932	(26.52, -28.88), 338	541	343	49.99	14.238	(24.72, -29.96), 526	839	332
GA 100 100	218.41	49.932 (49.932)	(26.52, -28.88), 338	541	342	49.99	14.238 (14.238)	(24.72, -29.96), 526	839	334

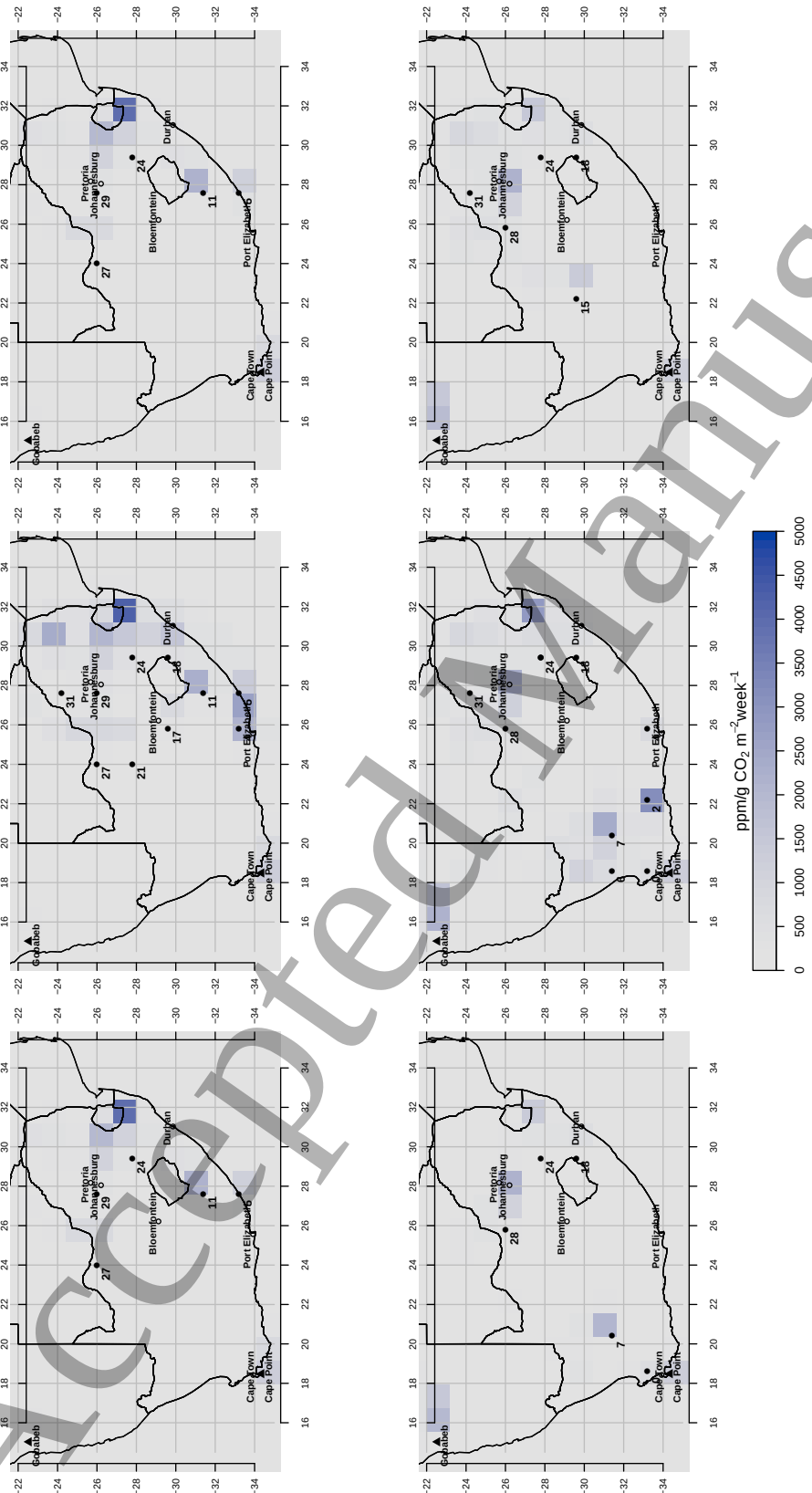

Figure 5: Cumulative sensitivities (ppm / g CO_2 m^{-2} week-1) over all sites under the solution with highest uncertainty reduction for each scenario, where pixels with large sensitivities indicate areas likely viewed by more than one station. The top row shows the sensitivities for the best solution for January under the original network design problem (left), established network scenario (middle), and subregion scenario (right). The bottom row shows the sensitivities for July.

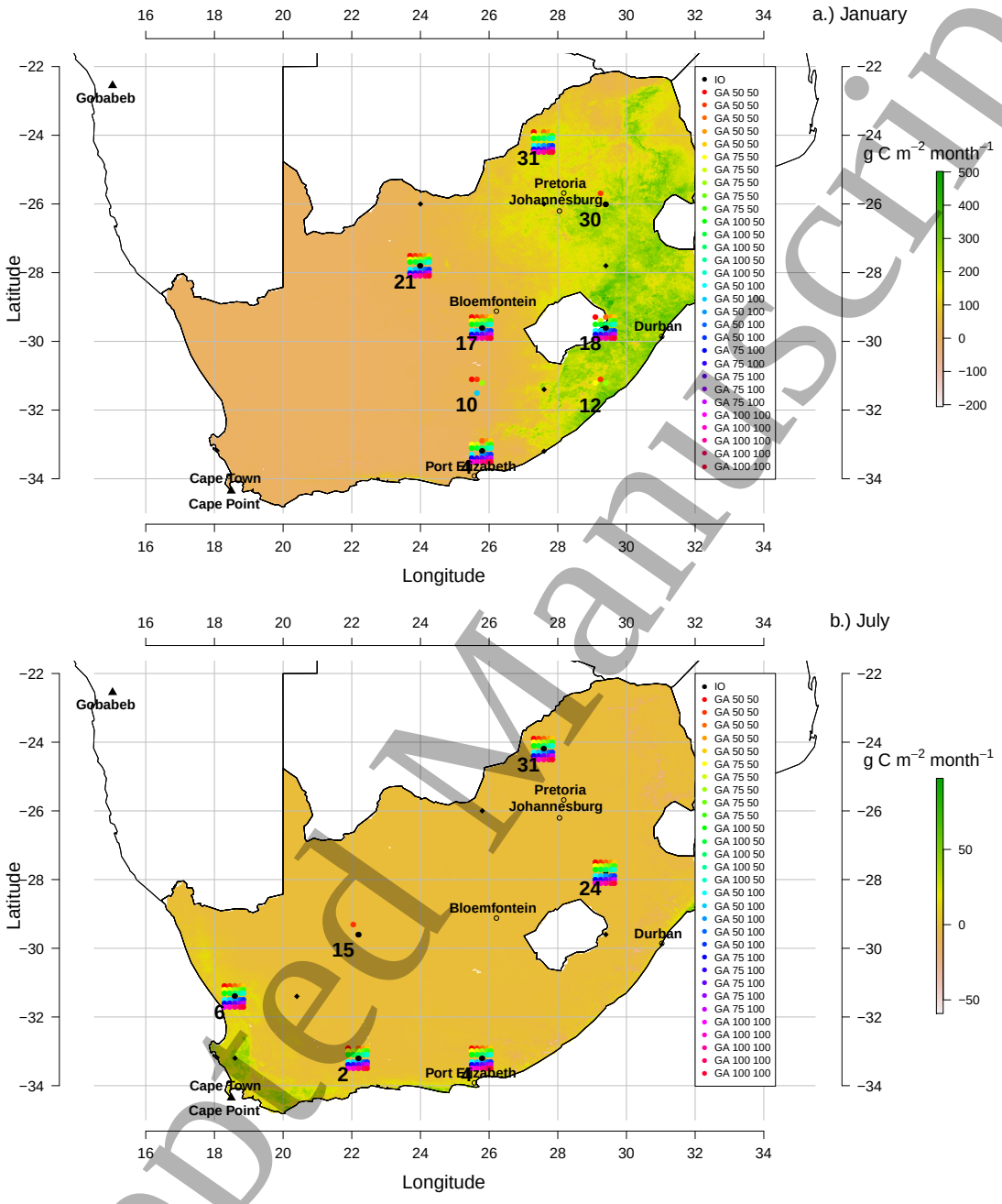

Figure 6: Optimal network locations for extended network case for (a) January and (b) July from each of the algorithm runs, overlaid on the net primary productivity (NPP) (g C m^{-2} month^{-1}). Code numbers for the GA solutions are the same as for Table 2. Numbers appearing on the maps are the station locations which have appeared in one or more network solutions. Black triangles - existing network stations of Cape Point and Gobabeb. Open circles - Major South African. Black closed circles - IO network solution. Coloured closed circles - GA solutions. Each colour represents a different GA solution. Points are laid out row by row, with the top row corresponding to the GA 50 50 and the bottom corresponding to the GA 100 100.

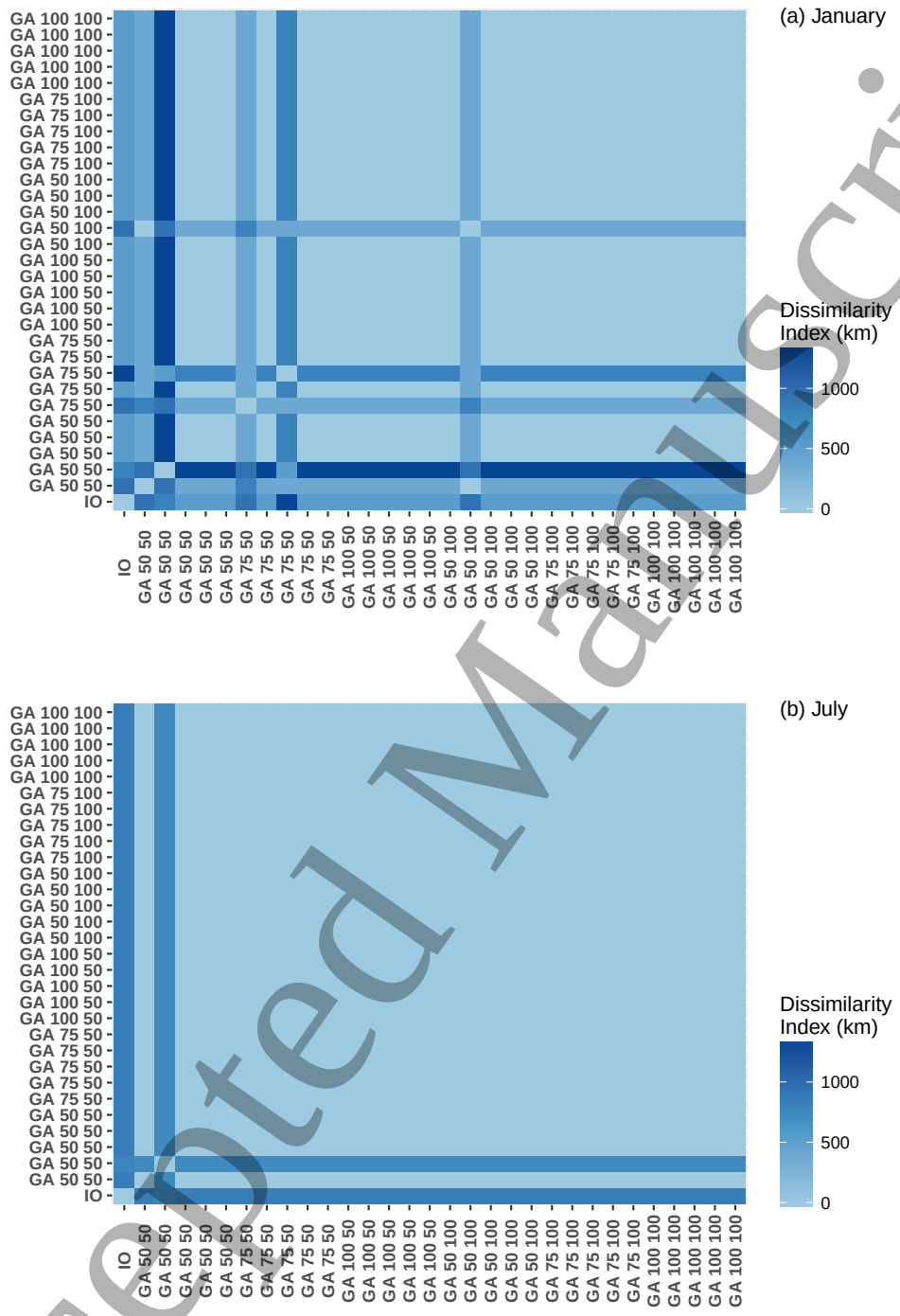

Figure 7: Heat maps displaying Dissimilarity Indices for the January (top) and July (bottom) optimal network solutions under the established network scenario, where the established network consists of the best five-member network from the original network design problem. The abbreviated names for the GA runs are the same as for Table 2.

3.3. Subregion Design Problem

For the subregion scenario, the uncertainty reduction was based only on the uncertainty within a subregion located on the eastern side of South Africa. The background stations, both located on the western side of the domain, would provide very little information about this region. The uncertainty reduction achieved for the January solutions ranged between 78.6% and 79.1%, and in July ranged between 51.1% and 51.4% (Table 3). The IO had the lowest uncertainty reduction for both the January and July solutions, although very close to the maximum uncertainty reduction achieved, differing by only 0.5% in January and 0.3% in July. In January, when there were both large fossil fuel sources and a large amount of biospheric activity in the subregion, the GA was always able to obtain a better solution compared with the IO. The consistency in the GA solutions for July was achieved with fewer evaluations compared with January.

For the subregion scenario, footprints of the measurement towers showed more overlap compared with both the original network design problem and the established network design problem, with the highest cumulative sensitivity occurring in January (Figure 5). This occurred in a pixel over the north eastern edge of South Africa, close to the Swaziland border, where the biospheric flux uncertainty was high. In all the best network solutions, across both January and July, the optimisation algorithms aimed to reduce the uncertainty over this region. By restricting the uncertainty optimisation over a subregion which contained this pixel, the network solutions could include multiple sites which viewed this location in order to have more comprehensive information about this source and therefore to produce a large uncertainty reduction overall.

The mean distances to the centroid for the January solutions were similar to those obtained for the original network design, with stations scattered throughout the subregion, particularly close to the major cities in the region (Figure 8). Two stations were outside of the subregion close to the borders of the subregion, one to the west of the subregion, near the city of Johannesburg and areas of high fossil fuel activity, and one to the south near regions of high biospheric activity. In July the clustering was lower, and stations tended to be located near the cities in the subregion, with one station located in the central part of South Africa, outside of the subregion. The DI values were higher in January and showed diversity in the solutions between the different optimisation runs, whereas for July the DI values were smaller and showed that there were only two solution sets that were obtained from the GA runs, where the best solution was obtained more reliably under either high number of iterations or high population size (Figure 9).

The run times for the optimisation procedures were very similar to those obtained for the original network design problem.

Table 3: Table of uncertainty reductions achieved by each network in the subregion network problem, together with the network's centroid point, the mean distance to centroid, dissimilarity index, and running times. Results are presented for the IO result, and results under the GA with 50, 75, and 100 iterations together with 50 population members (GA 50 50, GA 75 50, GA 100 50), and with 50, 75 and 100 iterations together with 100 population members (GA 50 100, GA 75 100, GA 100 100), where solutions were generated under the requirement that uncertainty needed to be reduced over the eastern subregion of South Africa

	January					July				
Algorithm Specification	Uncertainty kt C week^{-1}	Uncertainty reduction (%) (mean)	Centroid, mean distance to centroid (km)	DI with IO (km)	Run time (min)	Uncertainty kt C week^{-1}	Uncertainty reduction (%) (mean)	Centroid, mean distance to centroid (km)	DI with IO (km)	Run time (min)
Prior	2052.21					83.5				
IO	438.13	78.7	(27.24, -28.52), 300	0	14	40.86	51.0661	(27.24, -27.08), 284	0	16
GA 50 50	435.76	78.8	(27.24, -28.88), 326	891	68	40.56	51.4241	(26.88, -27.44), 335	533	76
GA 50 50	429.37	79.1	(27.24, -28.88), 357	726	67	40.56	51.4241	(26.88, -27.44), 335	533	75
GA 50 50	435.76	78.8	(27.24, -28.88), 326	891	68	40.86	51.0661	(27.24, -27.08), 284	0	75
GA 50 50	429.37	79.1	(27.24, -28.88), 17	726	67	40.56	51.4241	(26.88, -27.44), 335	533	75
GA 50 50	431.48	79 (78.96)	(27.24, -28.52), 11	1127	68	40.56	51.4241 (51.3525)	(26.88, -27.44), 335	533	75
GA 75 50	435.39	78.8	(27.6, -28.52), 24	360	100	40.56	51.4241	(26.88, -27.44), 335	533	109
GA 75 50	435.39	78.8	(27.6, -28.52), 24	360	99	40.86	51.0661	(27.24, -27.08), 284	0	108
GA 75 50	429.37	79.1	(27.24, -28.88), 17	726	100	40.56	51.4241	(26.88, -27.44), 335	533	109
GA 75 50	433.75	78.9	(27.6, -28.88), 17	1086	101	40.56	51.4241	(26.88, -27.44), 335	533	109
GA 75 50	429.37	79.1 (78.94)	(27.24, -28.88), 17	726	99	40.56	51.4241 (51.3525)	(26.88, -27.44), 335	533	110
GA 100 50	429.37	79.1	(27.24, -28.88), 17	726	129	40.56	51.4241	(26.88, -27.44), 335	533	144
GA 100 50	435.39	78.8	(27.6, -28.52), 24	360	130	40.86	51.0661	(27.24, -27.08), 284	0	146
GA 100 50	435.06	78.8	(27.6, -28.16), 17	755	133	40.56	51.4241	(26.88, -27.44), 335	533	145
GA 100 50	435.39	78.8	(27.6, -28.52), 24	360	132	40.56	51.4241	(26.88, -27.44), 335	533	147
GA 100 50	435.39	78.8 (78.86)	(27.6, -28.52), 24	360	133	40.56	51.4241 (51.3525)	(26.88, -27.44), 335	533	158
GA 50 100	429.37	79.1	(27.24, -28.88), 17	726	137	40.56	51.4241	(26.88, -27.44), 335	533	144
GA 50 100	429.37	79.1	(27.24, -28.88), 17	726	131	40.56	51.4241	(26.88, -27.44), 335	533	147
GA 50 100	435.39	78.8	(27.6, -28.52), 24	360	132	40.56	51.4241	(26.88, -27.44), 335	533	146
GA 50 100	429.37	79.1	(27.24, -28.88), 17	726	132	40.56	51.4241	(26.88, -27.44), 335	533	146
GA 50 100	429.37	79.1 (79.04)	(27.24, -28.88), 17	726	132	40.56	51.4241 (51.4241)	(26.88, -27.44), 335	533	145
GA 75 100	429.37	79.1	(27.24, -28.88), 17	726	194	40.56	51.4241	(26.88, -27.44), 335	533	221
GA 75 100	442.59	78.4	(27.24, -28.52), 30	1272	195	40.56	51.4241	(26.88, -27.44), 335	533	215
GA 75 100	429.37	79.1	(27.24, -28.88), 17	726	197	40.56	51.4241	(26.88, -27.44), 335	533	215
GA 75 100	435.39	78.8	(27.6, -28.52), 24	360	198	40.56	51.4241	(26.88, -27.44), 335	533	214
GA 75 100	429.37	79.1 (78.9)	(27.24, -28.88), 17	726	197	40.56	51.4241 (51.4241)	(26.88, -27.44), 335	533	214
GA 100 100	429.37	79.1	(27.24, -28.88), 17	726	256	40.56	51.4241	(26.88, -27.44), 335	533	282
GA 100 100	435.39	78.8	(27.6, -28.52), 24	360	255	40.56	51.4241	(26.88, -27.44), 335	533	284
GA 100 100	435.06	78.8	(27.6, -28.16), 17	755	263	40.56	51.4241	(26.88, -27.44), 335	533	287
GA 100 100	429.37	79.1	(27.24, -28.88), 17	726	254	40.56	51.4241	(26.88, -27.44), 335	533	285
GA 100 100	429.37	79.1 (78.98)	(27.24, -28.88), 17	726	258	40.56	51.4241 (51.4241)	(26.88, -27.44), 335	533	285

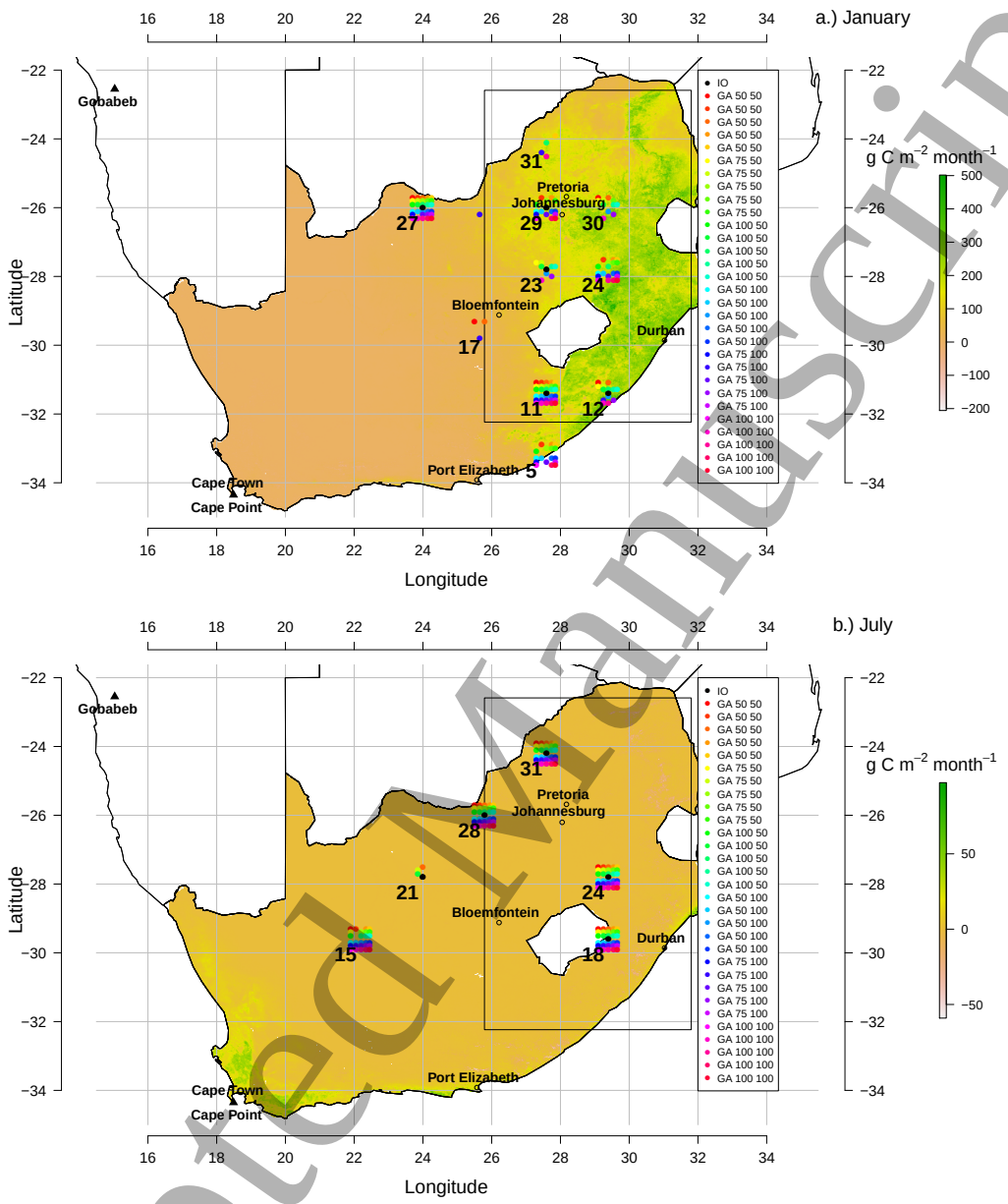

Figure 8: Optimal network locations for subregion case for (a) January and (b) July from each of the algorithm runs, overlaid on the net primary productivity (NPP) (g C m^{-2} month^{-1}). Subregion is represented by the grid box. Code numbers for the GA runs are the same as for Table 3. Numbers appearing on the maps are the station locations which have appeared in one or more network solutions. Major South African - open circles. Black triangles - existing network. Black closed circles - IO network solution. Coloured closed circles - GA solutions. Each colour represents a different GA solution. Points are laid out row by row, with the top row corresponding to the GA 50 50 and the bottom corresponding to the GA 100 100.

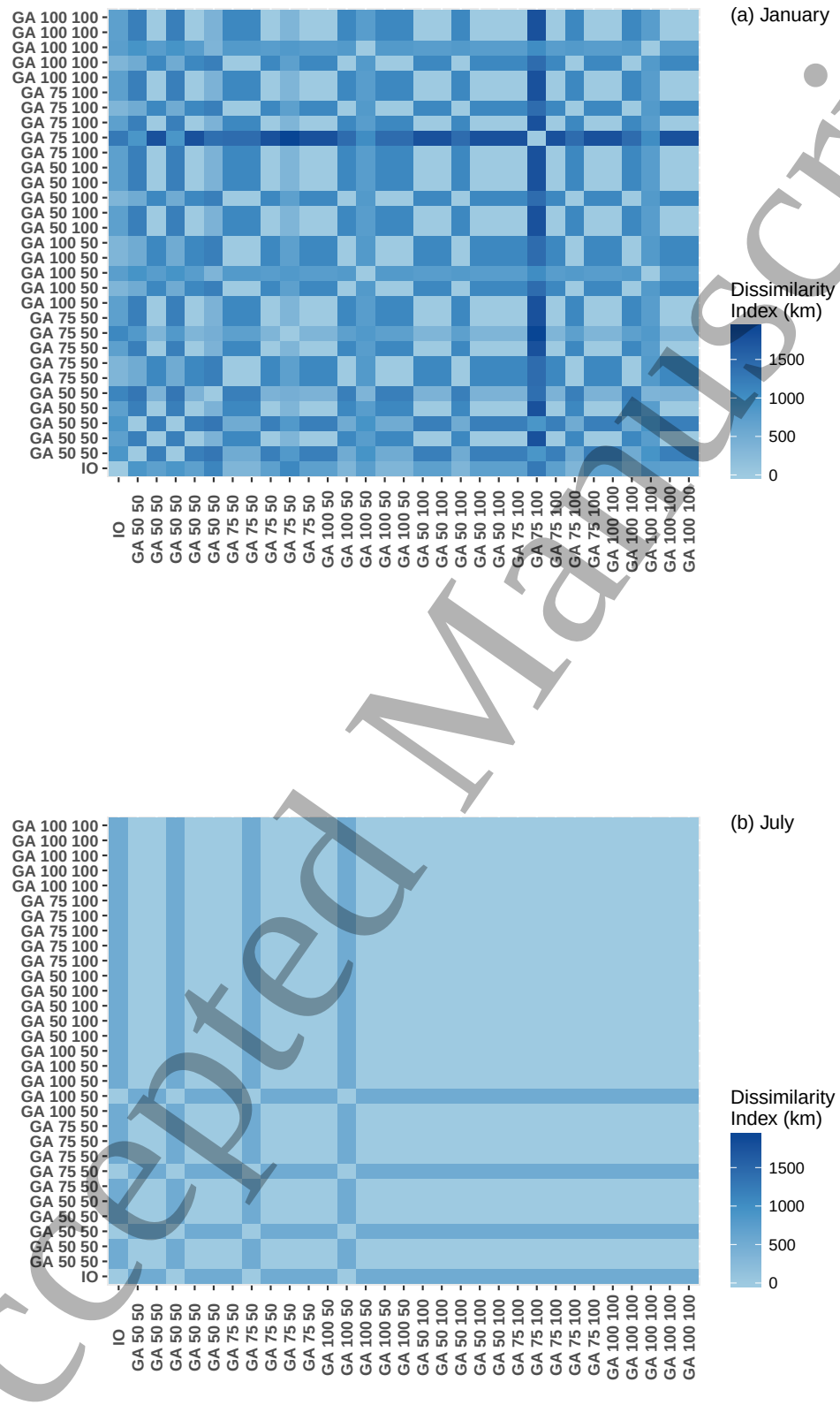

Figure 9: Heat maps displaying Dissimilarity Indices for the January (top) and July (bottom) optimal network solutions under the subregion scenario, requiring a solution to reduce the uncertainty in the eastern half of the country. The abbreviated names for the GA runs are the same as for Table 2.

4. Discussion and Conclusion

The results of this study show that although IO routine was not able to find the network with the global maximum uncertainty reduction, it found a network with uncertainty reduction only fractionally different from the best result obtained by means of the GA, at a much lower computational cost. Out of a five-member network solution, the IO differed by between one and two stations from the best GA solution. This occurred for the original network design problem, when solving for additional stations to an established network, and when solving for only a subregion of the domain. Moreover, the metrics for clustering of stations and dissimilarity showed that a very similar placement of stations would result. The advantage of the IO method over the GA method is that an evolution of results is obtained, which is useful for practical purposes. By identifying the station which on its own reduces the uncertainty by the most, it gives the decision makers the location of the site which should be prioritised. In addition, the running time for the IO is significantly shorter compared with the GA, which can play a role when computational resources are limited or when dealing with a larger or more complex domain, and where sensitivity tests need to be performed, for example using different configurations and estimates of the covariance matrices or using an alternative transport model. Therefore the IO algorithm is a viable alternative to the GA for optimal network design of atmospheric monitoring stations. In January, when the prior covariance matrix of the surface fluxes was more complex due to larger and more variable surface fluxes across a large proportion of the domain when compared with July, the difference in the uncertainty reduction between the best GA and IO solutions was also larger. In order to be confident of achieving the largest possible uncertainty reduction, the number of iterations and population size needed to be made large, and for the original network problem where the majority of the domain was unconstrained by the base network observations, there was still inconsistency between solutions in terms of placement. Fortunately, even if the GA settled on a sub-optimal solution, the fitness of this network was similar to the best available GA solution.

For the original network design problem, the GA was able to find the best solution, but not consistently, particularly for the month of January when the prior uncertainties of the fluxes were larger and more complex. In July, when the covariances were smaller and concentrated in small areas spread throughout the domain, the chances of the GA finding the best solution improved, resulting in greater consistency between runs. Increasing the number of iterations or population members did not guarantee the best solution, but uncertainty reductions were always higher or the same compared with the IO result. Therefore if very small reductions in the uncertainty reduction are worth pursuing, such as in the case where the existing network is already well established, the GA routine is preferable. Under these circumstances a large amount of resources, relative to the IO routine, would be required in order to run the GA with sufficient iterations and population members. These specifications depend on the complexity of the prior covariance matrix, with a larger number of iterations and population members

required to solve for networks under greater spatial variability in surface flux uncertainty. Therefore the number of iterations or population members should be tuned according to prior information available for a given network design problem.

We showed that even under more complicated scenarios, such as solving for additional stations to add to an established network or solving for the uncertainty in the fluxes of a subregion, the IO method was able to achieve a solution similar to the best GA solution, in terms of both location of stations and the uncertainty reduction achieved, but that the GA was always able to find a slightly better solution. The GA performed more consistently when the sources were concentrated within specific regions, such as around cities which occurred for the month of July. When the uncertainty resulting from the biospheric sources, which were more dispersed across the domain than the fossil fuel sources, had a dominant contribution to the overall uncertainty, such as for January in our case, the GA showed much more diversity in the network solutions, and a larger number of iterations and population members were required for convergence. The variability in the solutions was greater for the original network design problem than for the two additional scenarios, and the established network design for the month of July, where the gain in the uncertainty reduction was much lower and where the main sources of uncertainty were concentrated in a few places, showed the least amount of variability in the optimisation solutions.

The disadvantage of the GA procedure is that it does not supply the station which on its own results in the highest uncertainty reduction. This would require having to run the algorithm for a one station network. It also requires significantly more computational resources compared with the IO, without the guarantee of improvement over the IO. This implies that the user should not rely on a single run of the GA if the best solution is required, but rather have it run multiple times. Where computational resources are not limiting, the exact number of runs could be determined by the variability in the uncertainty reductions produced by multiple GA runs, where the number of evaluations of the GA can steadily be increased until the standard error of the mean uncertainty reduction under the set number of evaluations is below a required level. This would further increase the computation resources required to obtain the final solution, but would give the user alternative network solutions and an indication of the most important stations, as these would repeatedly appear in the solutions. With the gain in the uncertainty reduction so small for the GA best solution over the IO solution under all scenarios, it would suggest that it would be more worthwhile to invest in the improvement of the prior covariance estimates than in implementing a complex and resource intensive optimisation algorithm for this network design application. The network solution itself and the computational resources required for convergence of the GA is dependent on the complexity of the uncertainty covariance matrix of the prior surface fluxes. The network solution, regardless of optimisation algorithm, is only as good as the information provided for the uncertainties in the prior fluxes.

The GA and IO show that more than one good solution exists for the network design. Pragmatic considerations should be taken into account so that the resulting

network is feasible and cost effective. A network that is guaranteed to result in a reliable measurement record would be more valuable for constraining the overall flux uncertainty than one which has a slightly better assumed uncertainty reduction but more likely to have measurement gaps.

In this investigation we only considered the population size and number of iterations specified for the GA. We could also consider changing the probabilities assigned to cross-over and mutation. This could provide better ability of the GA solution to get away from local extrema. If we were to compare the specification of these parameters for the GA, we would need to ensure that the algorithms were compared under fair conditions (Črepinšek, Liu & Mernik 2014). If elitism is maintained, it would guarantee that the best solution always moves forward to the next iteration, which would ensure stability of the final solution. We recommend that the GA and its alternative parametrisations be used as part of a comprehensive sensitivity analysis when undertaking such an optimisation exercise.

We compared an evolutionary algorithm to a simple deterministic algorithm. An alternative deterministic algorithm to the IO is Decremental Optimisation (DO) (Curtis et al. 2004). DO starts with the maximal network, and eliminates sites from this network solution based on a fitness criterion, until the required network size is reached. Under the computational resources used for this analysis, DO would not have been possible as it would have resulted in a $\mathbf{H}$ approximately 7 times larger than for a five-member network, and a $\mathbf{C_c}$ that was 7^2 larger, which would have significantly increased the memory requirement for a single inversion result. This type of optimisation may be possible under a system with larger memory resources to allow the large matrix multiplications and matrix inversions required.

We assumed that the observation errors and flux uncertainties followed a Gaussian distribution. This is common practice in the field CO_2 flux inversions, and has made the technique described in this paper possible. This optimisation problem could be extended to relax the assumption of Gaussian errors.

5. Acknowledgements

Peter Rayner was in receipt of an Australian Professorial Fellowship (DP1096309). This worked was supported by parliamentary grant funding from the Council of Scientific and Industrial Research. The authors would like to thank the helpful commentary from Thomas Lauvaux on the implementation and post processing of the LPDM.

6. Appendix

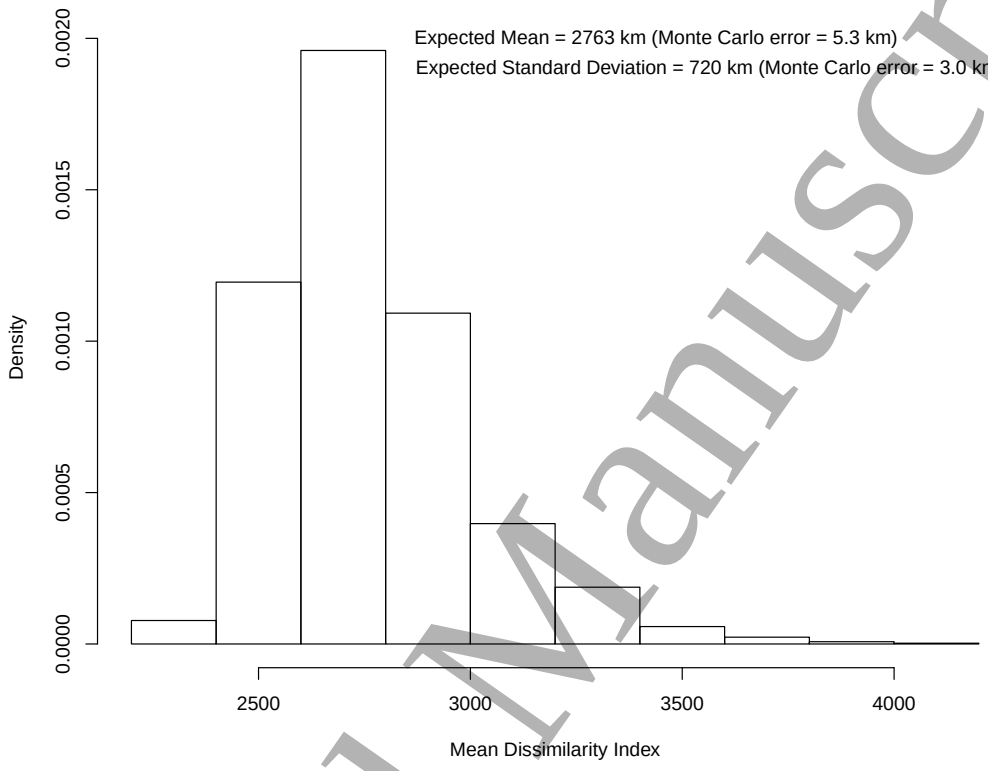

Figure 10: Histogram showing the distribution of the means for the dissimilarity index arising for 2000 generated pairs of network members. The expected mean and standard deviation for the dissimilarity index, together with Monte Carlo errors, are provided.

As the dissimilarity index (DI) was likely to have a non-normal distribution, the distribution characteristics were determined via simulation. The DI was calculated between every pair of randomly generated five-member networks within the domain, where 2000 of these five-member networks were randomly sampled from the set of available stations. For each simulation of 2000 network solutions, the mean, variance, minimum and maximum of the DI's was determined. These distribution characteristics should be invariant to the number of randomly generated solutions, which represents a large pool of potential network solutions. This was repeated 2000 times (the bootstrap sample size). The distribution of the mean DI's is plotted in Figure 9, and the expected mean and standard deviation, as determined from the empirical summary statistics of the bootstrap samples, together with Monte Carlo error estimates are provided.

References

Baker, D. F., Law, R. M., Gurney, K. R., Rayner, P., Peylin, P., Denning, A. S., Bourquet, P., Bruhwiler, L., Chen, Y., Ciais, P., Fung, I. Y., Heimann, M., John, J., Maki, T., Maksyutov, S., Masarie, K., Prather, M., Pak, B., Taguchi, S. & Zhu, Z. (2006), 'Transcom 3 inversion intercomparison: impact of transport model errors on the interannual variability of regional co_2 fluxes, 1988–2003', *Global Biogeochemical Cycles* **20**(1), GB1002.
URL: *http://dx.doi.org/10.1029/2004GB002439*

Berry, L. T. M., Murtagh, B. A., McMahon, G., Sugden, S. & Welling, L. (1999), 'An integrated galp approach to communication network design', *Telecommunication Systems* **12**, 265–280.
URL: *http://dx.doi.org/10.1023/A:1019102930443*

Bousquet, P., Ciais, P., Peylin, P., Ramonet, M. & Monfray, P. (1999), 'Inverse modeling of annual atmospheric co_2 sources and sinks: 1. method and control inversion', *Journal of Geophysical Research: Atmospheres* **104**(D21), 26161–26178.
URL: *http://dx.doi.org/10.1029/1999JD900342*

Bréon, F. M., Broquet, G., Puygrenier, V., Chevallier, F., Xueref-Remy, I., Ramonet, M., Dieudonné, E., Lopez, M., Schmidt, M., Perrussel, O. & Ciais, P. (2015), 'An attempt at estimating paris area co_2 emissions from atmospheric concentration measurements', *Atmospheric Chemistry and Physics* **15**(4), 1707–1724.
URL: *http://dx.doi.org/10.5194/acp-15-1707-2015*

Broquet, G., Chevallier, F., Bréon, F. M., Kadygrov, M., Apadula, F., Hammer, S., Haszpra, L., Meinhardt, F., Morgui, J. A., Necki, J., Piacentino, S., Ramonet, M., Schmidt, M., Thompson, R. L., Vermeulen, A. T., Yver, C. & Ciais, P. (2013), 'Regional inversion of co_2 ecosystem fluxes from atmospheric measurements: reliability of the uncertainty estimates', *Atmospheric Chemistry and Physics* **13**(17), 9039–9056.
URL: *http://dx.doi.org/10.5194/acp-13-9039-2013*

Chambers, L. (2001), *The Practical Handbook of Genetic Algorithms: Applications, 2nd Edition*, Chapman & Hall/CRC, USA.

Chevallier, F., Ciais, P., Conway, T. J., Aalto, T., Anderson, B. E., Bousquet, P., Brunke, E. G., Ciattaglia, L., Esaki, Y., Frohlich, M., Gomez, A., Gomez-Pelaez, A. J., Haszpra, L., Krummel, P. B., Langenfelds, R. L., Leuenberger, M., Machida, T., Maignan, F., Matsueda, H., Morgui, J. A., Mukai, H., Nakazawa, T., Peylin, P., Ramonet, M., Rivier, L., Sawa, Y., Schmidt, M., Steele, L. P., Vay, S. A., Vermeulen, A. T., Wofsy, S. & Worthy, D. (2010), 'Co_2 surface fluxes at grid point scale estimated from a global 21 year reanalysis of atmospheric measurements', *Journal of Geophysical Research: Atmospheres* **115**(D21), D21307.
URL: *http://dx.doi.org/10.1029/2010JD013887*

Ciais, P., Rayner, P., Chevallier, F., Bousquet, P., Logan, M., Peylin, P. & Ramonet, M.

(2010), 'Atmospheric inversions for estimating co_2 fluxes: methods and perspectives', *Climate Change* **103**(1-2), 69–92.
URL: *http://dx.doi.org/10.1007/s10584-010-9909-3*

Curtis, A., Michelini, A., Leslie, D. & Lomax, A. (2004), 'A deterministic algorithm for experimental design applied to tomographic and microseismic monitoring surveys', *Geophysical Journal International* **157**(2), 595–606.
URL: *http://dx.doi.org/10.1111/j.1365-246X.2004.02114.x*

Enting, I. G. (2002), *Inverse problems in atmospheric constituent transport*, Cambridge University Press, Cambridge.

Enting, I. G. & Mansbridge, J. V. (1991), 'Latitudinal distribution of sources and sinks of co_2: Results of an inversion study', *Tellus B* **43**(2), 156–170.
URL: *http://dx.doi.org/10.3402/tellusb.v43i2.15261*

Enting, I. G., Trudinger, C. M., & Francey, R. J. (1995), 'A synthesis inversion of the concentration and $\delta_1$3c of atmospheric co_2', *Tellus B* **47**, 35–52.

Gerbig, C., Lin, J. C., Wofsy, S. C., Daube, B. C., Andrews, A. E., Stephens, B. B., Bakwin, P. S. & Grainger, C. A. (2003), 'Toward constraining regional-scale fluxes of co_2 with atmospheric observations over a continent: 1. observed spatial variability from airborne platforms', *Journal of Geophysical Research: Atmospheres* **108**(D24), 4756.
URL: *http://dx.doi.org/10.1029/2002JD003018*

Grady, S. A., Hussaini, M. Y. & Abdullah, M. M. (2005), 'Placement of wind turbines using genetic algorithms', *Renewable Energy* **30**(2), 259–270.
URL: *http://dx.doi.org/10.1016/j.renene.2004.05.007*

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y., Ciais, P., Fan, S., Fung, I. Y., Gloor, M., Heimann, M., Higuchi, K., John, J., Kowalczyk, E., Maki, T., Maksyutov, S., Peylin, P., Prather, M., Pak, B. C., Sarmiento, J., Taguchi, S., Takahashi, T. & Yuen, C. (2003), 'Transcom 3 co_2 inversion intercomparison: 1. annual mean control results and sensitivity to transport and prior flux information', *Tellus B* **55**(2), 555–579.
URL: *http://dx.doi.org/10.1034/j.1600-0889.2003.00049.x*

Haber, E., Horesh, L. & Tenorio, L. (2008), 'Numerical methods for experimental design of large-scale linear ill-posed inverse problems', *Inverse Problems* **24**(5), 055012.
URL: *http://dx.doi.org/10.1088/0266-5611/24/5/055012*

Hardt, M. & Scherbaum, F. (1994), 'The design of optimum networks for aftershock recordings', *Geophysical Journal International* **117**(3), 716–726.
URL: *http://dx.doi.org/10.1111/j.1365-246X.1994.tb02464.x*

Hartline, J. & Sharp, A. (2006), An incremental model for combinatorial maximization problems, *in* C. Àlvarez & M. Serna, eds, 'Experimental Algorithms, 5th International Workshop, WEA 2006. Lecture Notes in Computer Science, vol 4007', Vol. 117, Springer, Berlin, Heidelberg, pp. 716–726.
URL: *https://link.springer.com/chapter/10.1007/11764298_4*

Jackson, D. D. (1979), 'The use of a priori data to resolve non-uniqueness in linear inversion', *Geophysical Journal of the Royal Astronomical Society* **57**(1), 137–157.
URL: *http://dx.doi.org/10.1111/j.1365-246X.1979.tb03777.x*

Jackson, D. D. & Matsu'ura, M. (1985), 'A bayesian approach to nonlinear inversion', *Journal of Geophysical Research: Solid Earth* **90**(B1), 581–591.
URL: *http://dx.doi.org/10.1029/JB090iB01p00581*

Kaminski, T. & Rayner, P. J. (2017), 'Reviews and syntheses: guiding the evolution of the observing system for the carbon cycle through quantitative network design', *Biogeosciences* **14**(20), 4755–4766.
URL: *http://dx.doi.org/10.5194/bg-14-4755-2017*

Kaminski, T., Rayner, P. J., Heimann, M. & Enting, I. G. (2001), 'On aggregation errors in atmospheric transport inversions', *Journal of Geophysical Research: Atmospheres* **14**(D5), 4705–4715.
URL: *http://dx.doi.org/10.1029/2000JD900581*

Lauvaux, T., Miles, N. L., Deng, A., Richardson, S. J., Cambaliza, M. O., Davis, K. J., Gaudet, B., Gurney, K. R., Huang, J., O'Keefe, D., Song, Y., Karion, A., Oda, T., Patarasuk, R., Razlivanov, I., Sarmiento, D., Shepson, P., Sweeney, C., Turnbull, J. & Wu, K. (2016), 'High-resolution atmospheric inversion of urban co_2 emissions during the dormant season of the indianapolis flux experiment (influx)', *Journal of Geophysical Research: Atmospheres* **121**(10), 5213–5236.
URL: *http://dx.doi.org/10.1002/2015JD024473*

Lauvaux, T., Uliasz, M., Sarrat, C., Chevallier, F., Bousquet, P., Lac, C., Davis, K. J., Ciais, P., Denning, A. S. & Rayner, P. J. (2008), 'Mesoscale inversion: first results from the ceres campaign with synthetic data', *Atmospheric Chemistry and Physics* **8**(13), 3459–3471.
URL: *http://dx.doi.org/10.5194/acp-8-3459-2008*

Nickless, A., Ziehn, T., Rayner, P. J., Scholes, R. J. & Engelbrecht, F. (2015), 'Greenhouse gas network design using backward lagrangian particle dispersion modelling part 2: Sensitivity analyses and south african test case', *Atmospheric Chemistry and Physics* **15**(4), 2051–2069.
URL: *http://dx.doi.org/10.5194/acp-15-2051-2015*

Patra, P. K. & Maksyutov, S. (2002), 'Incremental approach to the optimal network design for CO_2 surface source inversion', *Geophysical Research Letters* **29**(10), 97–1–97–4.
URL: *http://dx.doi.org/10.1029/2001GL013943*

Rayner, P. J. (2004), 'Optimizing CO_2 observing networks in the presence of model error: Results from Transcom 3', *Atmospheric Chemistry and Physics* **4**(2), 413–421.

Rayner, P. J., Enting, I. G. & Trudinger, C. M. (1996), 'Optimizing the CO_2 observing network for constraining sources and sinks', *Tellus B* **48**(4), 433–444.
URL: *http://dx.doi.org/10.1034/j.1600-0889.1996.t01-3-00003.x*

Rayner, P. J., Raupach, M. R., Paget, M., Peylin, P. & Koffi, E. (2010), 'A new global gridded data set of CO_2 emissions from fossil fuel combustion: Methodology and evaluation', *Journal of Geophysical Research: Atmospheres* **115**(D19), D19306.
URL: *http://dx.doi.org/10.1029/2009JD013439*

Rodgers, C. D. (2011), *Optimal Linear Inverse Methods*, World Scientific, Hackensack, N. J.
URL: *http://dx.doi.org/10.1142/9789812813718_0004*

Seibert, P. & Frank, A. (2004), 'Source-receptor matrix calculation with a lagrangian particle dispersion model in backward mode', *Atmospheric Chemistry and Physics* **4**(1), 51–63.
URL: *http://dx.doi.org/10.5194/acp-4-51-2004*

Tarantola, A. (1987), *Inverse Problem Theory*, Elsevier, New York.

Tarantola, A. (2005), *Inverse problem theory and methods for model parameter estimation*, Society for Industrial and Applied Mathematics, Philadelphia.
URL: *http://dx.doi.org/10.1137/1.9780898717921*

Uliasz, M. (1994), Lagrangian particle modeling in mesoscale applications, *in* P. Zannetti, ed., 'Environmental Modelling II', Computational Mechanics Publications, Southampton, pp. 71–102.
ΩČrepinšek et al.

Črepinšek, M., Liu, S.-H. & Mernik, M. (2014), 'Replication and comparison of computational experiments in applied evolutionary computing: Common pitfalls and guidelines to avoid them', *Applied Soft Computing* **19**, 161–170.
URL: *http://dx.doi.org/10.1016/j.asoc.2014.02.009*

Wu, L., Bocquet, M., Chevallier, F., Lauvaux, T. & Davis, K. (2013), 'Hyperparameter estimation for uncertainty quantification in mesoscale carbon dioxide inversions', *Tellus B* **65**(1), 20894.
URL: *http://dx.doi.org/10.3402/tellusb.v65i0.20894*

Ziehn, T., Nickless, A., Rayner, P. J., Law, R. M., Roff, G. & Fraser, P. (2014), 'Greenhouse gas network design using backward Lagrangian particle dispersion modelling part 1: Methodology and australian test case', *Atmospheric Chemistry and Physics* **14**(17), 9363–9378.
URL: *http://dx.doi.org/10.5194/acp-14-9363-2014*